4Q－6

The Worldwide Multilingual Computing (6): Multilingual Text Interprocess Communication and Input Mechanism

Hidejiro Daikokuya†, Yutaka Kataoka*, Kazutomo Uezono†, Tomoko Kataoka*, Tadao Tanaka‡, Toshio Oya†, Kenji Maruyama‡, Shoichiro Yamanishi† and Hiroyoshi Ohara†

* Centre for Informatics, Waseda University † School of Science and Engineering, Waseda University
‡ Research and Development, Japan Computer Corporation

1. Introduction

Nowadays, the computers are physically connected together by worldwide computer networkings. However, ensuring internationalized *Interprocess Communication* (IPC) has not been considered enough.

To establish international communication among computers, ISO 2022 [1] specifies character codeset extension methods – encoding schemes – which only determines paths to allocate GCSs (*Graphic Character Sets*) and CCSs (*Control Character Sets*) to *In-Use Table* via *Intermediate Tables* (*G0, G1, G2, G3*). Needless to say, the all encoding schemes in ISO 2022 must be supported for the international communication.

However, one codepoint in a character set which is specified by ISO 2022 does not stand for one *Character* [Talk 1, 4]. The number of the codepoints in a GCS can be extended by combinations of the codepoints in the GCS described in the GCS's specification [2, 3] – codepoint extension, or in ISO 6429 [4], which was mainly designed for output devices. To establish international communications which require presentations of characters, information determining such extended codepoints and specifying presentation – direction of words, direction of line progress and final glyph must be sent even if the other side is not an output device.

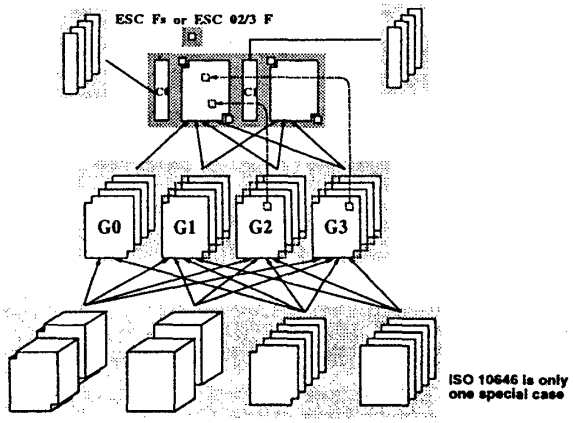
An *Input Method* (IM) which assists complicated character inputs for applications as client processes must use the information described above. Thus, multilingual interprocess communication (ML IPC) should have functions accepting all encoding schemes in ISO 2022 and presentation of characters after determining one character by extensions in a character set. Therefore, to realize ML IPC, developing multilingual IM makes essential information and functions clear.

2. Requirements of Multilingual Interprocess Communication

The ML IPC should manage two levels of communication, 1) ISO 2022 and 2) presentation after determining a character. When all senders and all receivers share the same *Meta Converter System* and when all take the *Global IOTMC Model* [2], both layers of communication can be performed without any inconsistency.

In communication by ISO 2022, a sender and a receiver should accept all encoding schemes and the same set of GCSs and CCSs with keeping states by invocation and designation sequences during multiple sites communication. It is possible to keep status by preparing a pointer to data to retain the state for each communication. But ISO 2022 does not specify methods to inform a 'signal' to a sender when it has received an unknown terminal character. Thus, to perform a complete communication, higher level protocol that informs available invocation sequences, designation sequences terminated by *terminal byte* with GCS/CCS name each other before starting communication – *Character Set Negotiation* (CSN).

After CSN, both sender and receiver can decide their actions by their policies, i.e., quitting communication or starting communication within the GSC/CCS limitation or ignoring the limitation. When starting communication within the limitation, senders should filter character sets and select invocation sequences for receivers. When ignoring the limitation, receivers should filter character sets and encodings from senders. The filtering problems were solved by *Interprocess Communication Assisting Functions* (ICAF) and *GCS/CCS Information Functions*, which were discussed in Talk 5. Note that IS 13194 has specific designation sequences to select one of 11 scripts. Thus, the extension schemes including IS 13194 beyond ISO 2022 are shown in Figure 1.



The Total Extension Methods beyond ISO 2022

Solving the problems above is necessary to keep *data exchangeability*.

Presentation information to specify a glyph of a character should be transferred for an application interactively specifying characters. By the researches of characters and codepoints in codesets, relation of a glyph and a character mapped from codepoints was discovered [Talk 1]. Thus, to specify a glyph, a character must be determined from a codepoint sequence, i.e., a codepoint extension rules that specify characters from codepoint

sequence must be the same among senders and receivers. Using the same codepoint extension rule is just essential especially for *Conjunct Syllabic* like *Devanagri scripts* and *Thai scripts* [5]. Since the rules are quite large, it is inappropriate way for senders and receivers to transfer the rules. Thus, before communication gets started, the rules must be distributed, i.e., a system that can be programmed by the rules can satisfy ML IPC [2]. Note that multiple codepoint extension rules for one character set may be possible. Therefore, a recognition and a selection of the codepoint extension rules among senders and receivers are required. In order to keep computability, also higher level protocol must satisfy the requirements above.

The presentation information were extracted from the information discovered by the researches to determine one glyph from a character [2, 6]. The presentation information consists of origin of line, direction of line progress and current direction for direction dependent characters – level 1. Also character direction, selection of forms of *Perso-Arabic* [2], direction of word progress and script angle may be specified – level 2. When the Meta Converter System is used, only level 1 information is essential and level 2 are needed for specific control, because the Meta Converter System can supply level 2 information as default. On the other hand, if a receiver does not have the Meta Converter System, all level 2 information must be required by a sender to control the receiver all the time.

The information of level 1 and 2 can be presented by the specification described in ISO 6429. But since ISO 6429 is not appropriate for the purpose, another international standard should be defined. The Meta Converter System is programmable and any control sequences corresponding to the information may be set. Higher level protocols were designed by the researches and a mechanism called *ML IPC model* was developed as a satellite model of the *Global I/O TM/C* System.

3. Multilingual Input Method based on ML IPC

When implementing a multilingual IM based on ML IPC, Input Method Protocol (IMP) is required. The IMP must satisfy the requirements solved by the higher level protocols as well as both level 1 and level 2 informations. But current IMP for X11 R5 (partial multilingual) and R6 (locale based) could not satisfy the requirements, level 1 (partially covered) and level 2. Thus, the protocol cannot support vertical drawing and right-to-left drawing. As a special case of requirement for IMP, delimiting a character by specific signal is required. TIS 620-2533 needs a delimiter to determine a character but the codeset does not contain any appropriate delimiter. Thus, if required during communication, IPC specific higher level protocol that delimits a character must be set over ISO 2022 level communication.

The generalization functions and internal structure of Multilingual Input Methods were already done by researches based on the characters and orthographies [7]. By our implementation trial of multilingual IM based on ML IPC for non-native speakers, other effective functions could be generalized. Adding a function that selects a character in a dictionary by keystrokes and all matched characters before keystrokes are complete, candidates of unknown complete keystroke characters can be found. This new function and other generalized functions were integrated into new Multilingual IM based on ML IPC. The Keyboard assignment of characters must be shown on the display by IM for the non-native speakers when codeset is changed.

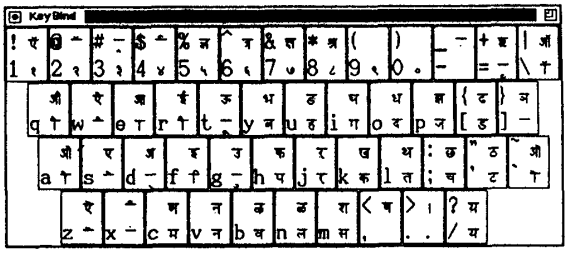


Figure 2. Example of Keyboard Assignment

The new IM is highly programmable by internal interpreters and also it can customizable according to users' needs.

4. Summary

The essential information and functions establishing multilingual interprocess communication after physical connection is engaged were discovered. And essentials for multilingual IM protocol were also found. By implementation using those above, multilingual IM which has more generalized GUI was developed. Non-native speakers can get input easily by the helps of the GUI. Thus, the IM can be widely used for language education and library database and so other requirements to be multilingual as well as daily use.

References

[1] ISO/IEC 2022: 1986, Information processing – 7-bit and 8-bit coded character sets – Code extension techniques.

[2] Kataoka, Y., et al., 1995. Codeset Independent Full Multilingual Operating System: Principles, Models and Optimal Architecture, IPSJ SIG System Software & Operating System, 68-4, pp 25-32.

[3] IS 13194: 1991, Indian Script Code for Infromation Interchange – ISCII, Bureau of Indian Standards, India

[4] ISO/IEC 6429: 1992, Information processing – Control functions for coded character sets.

[5] TIS 620-2533 (1990), Thai Character Codes for Computers, Thai Industrial Standards Institute, Ministry of Industry, Thailand.

[6] Uezono, K. et al., Multilingual I/O and Text Manipulation System (2): The Structure of the Output Method Drawing the World's Writing Scripts beyond ISO 2022, Proceedings of the 49th General Meeting of IPSJ, Vol. 3, September 1994, pp 301-302.

[7] Kataoka, Y. et al., A model for Input and Output of Multilingual text in a windowing environment, ACM Transactions on Information Systems, Vol. 10, No. 4, October 1992, pp 438-451.