The Worldwide Multilingual Computing (2):

4Q－2

Functions, Model, Design and Architecture of Multilingual I/O TM/C System

Kazutomo Uezono†, Yutaka Kataoka*, Tomoko Kataoka*, Tadao Tanaka‡, Toshio Oya†, Hidejiro Daikokuya†, 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

By definitions of *Character* and relations among character, codesets and glyphs, functions to be realized were defined [4, Talks 1, 3]. In order to keep backward compatibilities and to satisfy all requirements, *Meta Converter System* based on *Global IOTMC Model* was established. By the Model and the System, total consistencies can be ensured. The Meta Converter System – kernel of the environment – can provide all information required and can convert codesets and encoding schemes. X Window System and other libraries were converted by the System to be multilingual.

2. Definition of Multilingual I/O TM/C System

A Multilingual I/O TM/C System should have the following four components: *Input*, *Output*, *Text Manipulation* and *Interprocess Communication* [4]. To realize mixing all code sets independent from any orthography of a language, each component in the system must avoid hard-coding. Therefore, the following two requirements must be satisfied by the system: 1) the system should have data files that describe definitions of all information for the four components, and 2) only one executable code of each component of the four should provide informations and perform functions for all process. A system satisfying above requirements qualifies as a generalized and unified system.

3. Models for Multilingual Computing

The *Global IOTMC Model* does not have Locale tables but has one *Global table*, in which all GCSs and *Control Character Sets* (CCSs) are registered (Fig. 1).

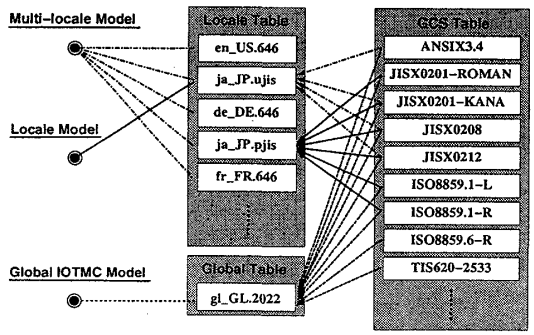

Figure 1. The Models for Multilingual Computing

Once calling *InitGlobal* function to initialize, all graphic character sets and control character sets can be used. To realize Text Manipulation for any purpose, this model provides the *Text Manipulation Functions*. To support Interprocess Communication, *Interprocess Communication Assisting Functions* always run under this model.

The *Multi-Locale Model* provides a set of *Multi-code-set* (MCS) locales by the *OpenLocale* function that returns each Locale-ID. By using of the ID, MCS can get I/O simultaneously, but this model is a subset of the Global IOTMC Model. POSIX Locale Model [1, 2] is a subset of the Multi-Locale Model.

4. Definitions of the Relations among mb, WC, TMC and Final Glyph Sets

Mb is a set of GCSs and CCSs. Since a mb codepoint does not always stand for a *Final Glyph* nor a *Control Function*, mb codepoints are *Non-Fixed-Length* (NFL) codepoints. The mb codepoints are extended to *Extended Codesets* by ISO 2022 [5] and combination of codepoints in GCSs and CCSs for codepoint extensions (ISO 6429 [6], IS 13194 [7], etc.).

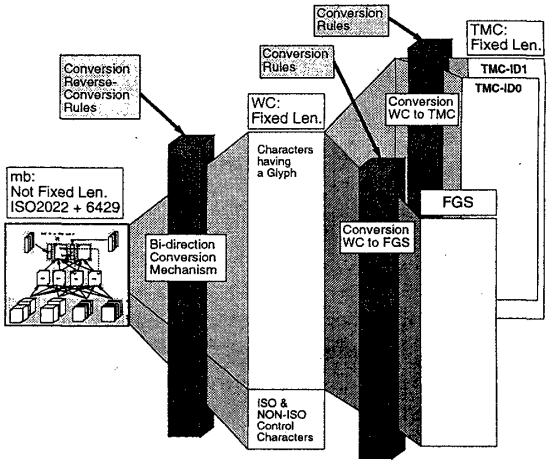

Figure 2. Relations among mb, WC, TMC and FGS

The Extended Codesets are mapped to WC. A WC codepoint must satisfy the following requirements: 1) to have information for presentation, 2) to have information for basic text manipulation (insertion, deletion, etc.) [Talk 4], 3) to keep codeset/language independency, and 4) to ensure reverse-conversion. To satisfy the above, WC is normalized to Character [Talk 1]. The set is the largest and involves all informations of GCSs and CCSs. WC is unique in a system to keep consistency among all models. WC can be used for text manipulation without

codeset/language dependency. To manipulate text codeset/language dependently or glyph dependently, *Text Manipulation Code* (TMC) is defined. TMC is converted from WC with necessary informations and/or glyph information given by OM. *Final Glyph Sets* (FGSs) are mapped from WC.

The overall relations among mb, WC, TMC and FGS are shown in Figure 2.

5. Design of the Meta Converter System

Meta Converter System developed optimally based on Global IOTMC Model. The system consists of *Meta Converter Table Compiler* (MCTC), *Trans-Unit Converter, GCS/CCS Information Functions, Character Information Functions, Text Manipulation Functions* and *Interprocess Communication Assisting Functions* (ICAF). Trans-Unit Converter converts to/from mb/WC/TMCs by finite number of essential functions (Concatenation, Conjunct_form, etc.). WC has 'hints' in itself to generate original mb encoding scheme [8]. But all schemes are changeable in pointers to states given from callers.

All data for conversions are stored in layered files. MCTC compiles the files and generate optimized automaton bodies. The bodies are loaded into Meta Converter system and are shared by all processes. Note that the bodies are mapped into virtual memory space. Thus, no hard-coding and single structure is ensured to keep the overall consistency. GCS/CCS Information Functions and Character Information Functions extract information from GCS/CCS and WC/TMC, respectively. Text Manipulation Functions manipulate WC/TMC strings. ICAF can convert encoding schemes to others to ensure encoding scheme independent interprocess communication.

6. Architecture of the Multilingual I/O TM/C System

Multilingual I/O TM/C System is constructed by the Meta Converter System (Fig. 3). *Multilingual OM* specifies Final Glyph from WC with glyph assignment rules [9]. Note that the character order in display image differs from that in memory image. *Multilingual IM* [Talk 6] communicates with clients through *Encoding Converter* that absorbs different encoding schemes and states.

All functions related to Locale, mb and WC were newly developed by using of the Meta Converter System and OS's libraries were replaced by the functions. WC/TMC Text Manipulation function set was prepared in IOTMC library. With OM and IM, a set of the generalized text manipulation functions for WC, was provided as *Multilingual Text Widget* on X Window System [Talk 5]. The widget supports formatting and editing all characters in the world correctly.

Interprocess communication requires not only satisfying ISO 2022 but also displaying presentation information. Thus the ICAF returns such information as any sequence by any request.

Note that multilingual FORTH is a part of the system [Talk 7].

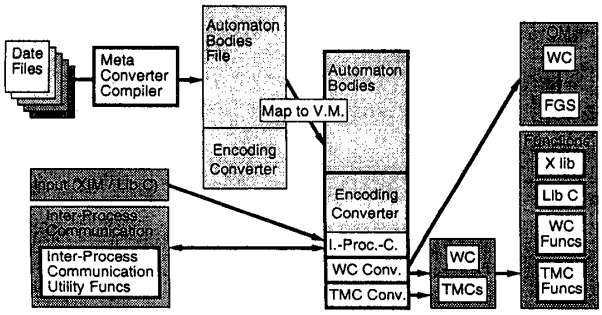

Figure 3. Multilingual I/O TM/C System

7. Summary

Our Multilingual researches could define essential necessities of multilingual computing environment. And the necessities are much larger than current national and international specifications – i.e., all essential information for multilingual computing was discovered. By the Multilingual I/O TM/C System optimally designed, a multilingual computing environment was provided. The most important result is that processing highly codeset/language dependencies and/or personal requirements can be satisfied by Multilingual programming language using TMC [Talk 4].

The System can easily provide multilingual tools and applications, and Multilingual Operation System could be developed [Talk 8].

References

[1] ISO/IEC 9899: 1990, Programming language C.

[2] ISO/IEC 9945-1: 1990, Information technology – Portable Operating System Interface (POSIX) Part 1: System Application Program Interface (API) [C Language].

[3] Kataoka, Y., et al., Multilingual I/O and Text Manipulation System (1): The Total Design of the Generalized System based on the World's Writing Scripts and Code Sets, Proceedings of the 49th General Meeting of IPSJ, Vol. 3, September 1994, pp 299-300.

[4] 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.

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

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

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

[8] Tanaka, T., et al., Multilingual I/O and Text Manipulation System (4): The Optimal Data Format Converter to/from MB/WC/TMC, Proceedings of the 49th General Meeting of IPSJ, Vol. 3, September 1994, pp 305-306.

[9] 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.