Kanji Characters in Japan – Remaining Challenges

Toshihiro Enami

Fujitsu Research Institute
16-1, Kaigan 1-chome, Minato-ku, Tokyo 105-0022, Japan
enami.toshihiro@jp.fujitsu.com

Abstract. The Japanese Government has set its sights on becoming number one in the world in ICT, as seen in the announcement of an e-Japan strategy by the IT strategy organization established in 2000. However, according to the United Nations E-Government Development Index, Japan's ranking has remained low despite its progress in information infrastructure. The reason for this is that the Japanese government did not integrate the code and standardize the data which are needed to use ICT across the whole country. The government introduced the national ID bill into the Diet last year, but the issue of Kanji characters, i.e., how to define the Japanese Kanji character set, remains unresolved because this issue, especially as it related to Kanji characters of names, includes a complex problem of interface between human and machine. I think the current proposed solution will not be successful because it ignores the issue of human interface. I insist that the Kanji character issue should be viewed from the perspective of human interface, and I propose a solution whereby the government should regulate by law the range of Kanji characters used by ICT, rather than increasing the number of Kanji character used.

Keywords: e-Government, Character Code, Kanji Character.

1 Introduction

In recent years, countries around the world have introduced e-Government and it continues to spread its roots. EGDI (2012) shows the e-Government ranking of 190 countries, with an additional 3 countries having no online services at all. The Japanese government passed an ICT law in 2000, and the next year it announced that it would aim to become number one in ICT through its e-Japan strategy. Afterwards, high speed information infrastructure was built, but e-Government did not progress in terms of using ICT. Japan was ranked 11th in 2008, 17th in 2010, and 18th in 2011.

By contrast, Korea was number one in 2010 and 2011. The difference between Japan and Korea is increasing. Shimada, Enami, and Yoshida (2012) discuss this difference from many perspectives and conclude that Japan is caught in a vicious cycle and Korea in a virtuous one. The most critical point is that the Japanese government did not recognize the importance of building a legal system to optimize the whole country through ICT, especially by destroying the bureaucratic silos.

S. Yamamoto (Ed.): HIMI/HCII 2013, Part I, LNCS 8016, pp. 20–29, 2013.

In 2012, the My Number (national ID) bill, which would build a legal system to identify Japanese people, was introduced into the Diet. But Kanji character code, as another important legal system, remains unpromising. On the surface, the issue of Kanji characters seems to be a problem between technology and culture, but it fundamentally includes the issue of interface between human and machine. We must confront this issue from this point of view.

2 Methodology

I adopt the following methodology for this study.

2.1 The History of Japanese Character Code and the Issue of Kanji Characters in Names

I begin with a survey of the history of Japanese character code as mapped onto computer systems, followed by a look at its present state. I then describe the issues of Kanji characters in the field of administrative procedure and how the government would address this issue based on the reports of METI (Ministry of Economy, Trade and Industry)[1] and MIC (Ministry of Internal Affairs and Communications)[2] .

I identify 2 perspectives which the government overlooks in order to find a path towards solving this issue. I analyze the issue from these perspectives, as outlined in 2.2 and 2.3.

2.2 The Cost Perspective (People's Burden)

One perspective is that of cost. Japan uses different types of character sets, or Gaiji, which are defined broadly as a character not included in a character set and narrowly as a character outside of JIS level-1 and level-2. I will calculate the economic loss which results from using these in the field of administrative procedure and in the private sector. The Japanese people must cover any economic loss through taxes or the high cost of products, and we should discuss whether it is right or not to make the people pay for this loss.

2.3 The Recognition Perspective

Another perspective is that of recognition. I present data on the recognition of Kanji characters by Japanese people and show the big difference in recognition (speed and accuracy) between a standard range (JIS level-1 and level-2) and a broad range (includes Gaiji) of characters.

In addition, I would like to consider the recognition of seniors and foreigners because the population of these demographics is increasing in Japan.

---

[1] METI/IPA(2011).

[2] Fuji Xerox Co., Ltd.(2012).

2.4 The Proposal of a New Solution

I propose a new solution, as discussed above. In particular, I point out that the second perspective includes the issue of interface between human and machine, and therefore technological solutions would not work. I propose a solution using the legal system, rather than the technological solutions using UTF-16 and IVS (Ideographic Variation Sequence)/ IVD (Ideographic Variation Database) proposed by the government.

3 The History of Japanese Character Code and the Issue of Kanji Characters in Names

Let us first look at the history of Japanese character code as mapped onto computer systems. Japanese has 3 types of character: Kanji are based on Chinese characters; Hiragana are derived from certain Kanji; and Katakana are derived from a different set of Kanji. Hiragana is a syllabary which complements Kanji in expressing Japanese, and Katakana is a syllabary used to express imported words and mimetic/onomatopoeic words. Each syllabary includes about 50 characters and the set of Kanji includes more than 50 thousand.

From early on, Japan needed multi-byte character code systems mapped into its computer systems because Japanese must use these many characters. In Japan, JIS X0201 was defined based on ISO646 as a 1-byte code system. And JIS X0208, JIS X0212, and JIS X0213 are defined based on ISO2022 as multi-byte code systems. JIS X0208 was defined in 1978 and includes about 6000 Kanji characters (JIS level-1 and level-2). JIS X0213 was defined subsequently in 2000 and includes about 3700 additional Kanji characters (JIS level-3 and level-4). There are also Japanese character code systems which integrate character set and encoding scheme: EUC-JP for UNIX and Shift-JIS for PC. Mainframe computers have different proprietary Japanese character code systems depending on the vendor.

Later on, Unicode (ISO10646) was proposed as a universal character code system. This system adopts a separation of character set and encoding scheme. The popular Unicode standards are UCS-2, UCS-4, UTF-16, UTF-8, and UTF-32. Character set and encoding scheme are integrated in UCS, but UTF is an encoding scheme only. These standards include from 65,000 to 2 billion code points, which means that all the characters in the world can be used in this system. More importantly, this system can use a very broad range of characters, including old or historical characters for academic research.

In the field of Japanese administrative procedure, fewer than 10,000 characters (mainly JIS level-1 and level-2) have been used since mainframe computers could use Japanese characters. But the developing Koseki[3] system of the late 90's and the Juki[4] system of the early 2000's needed more characters. The Koseki character set was defined with about 56,000 characters and Juki with 21,000. In fact, not all characters

[3] Census registration system.
[4] Resident information system.

used in names were defined; undefined characters are managed using image data in the Juki system and with paper in the Koseki system.

The relationship among the JIS, Juki, and Koseki character sets is shown in Figure 1. These characters are used in the field of administrative procedure, mainly in names. This figure shows that the total number of characters is around 60,000, but each character cannot be identified correctly between character sets because each system has different definition rules, which are the way of defining differences between form and design in characters. The essential reason for this disparity is the inter-ministry bureaucratic silos. JIS is the jurisdiction of METI, Juki is that of MIC, and Koseki is that of MOJ (Ministry Of Justice). In addition, the Japanese language, including Kanji characters, is the jurisdiction of MEXT (Ministry of Education, Culture, Sports, Science & Technology).

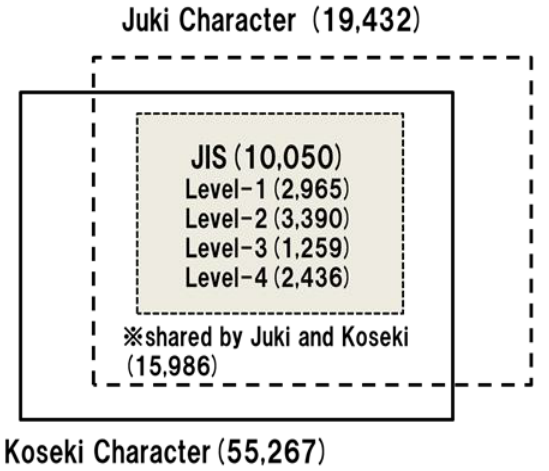

Fig. 1. Overview of JIS, Juki, and Koseki (based on data from METI/IPA(2011))

As the interconnectivity of society progresses, problems are occurring due to the current state of Kanji characters. Data exchange between government computers and other computers is not going well, data cannot be shared between public and private sectors or even among public sectors, and changing systems will be very costly.

Based on its "Report on Character Integration Infrastructure Project"[5] , METI integrated the 3 character sets to solve these problems and published the integrated character table (MJ character table). This report doesn't have a clear conclusion, but it seems to say that this problem can be solved in the future by identifying Gaiji with the MJ character table and implementing UTF-16 and IVS, as it says: "we are going to

[5] METI/IPA(2011).

use about 60,000 characters by encouraging the adoption of UTF-16 and IVS on top of the foundation of the MJ character table"[6] .

However, I do not think the adoption of UTF-16 and IVS will solve the following challenges.

- Japan has no integrated, comprehensive set of rules, and no one has the authority to decide on and register new characters for IVS/IVD.
- Currently, there are about 60,000 characters which cannot be used on standard PCs, and the public electronic signature system accepts alternative characters instead of Gaiji. Will people accept these 60,000 characters?
- Even if standard PCs could use these 60,000 characters, can we (citizens and government) use them correctly in the field of administrative procedure?

In addition, the "Report on Gaiji used by Municipalities"[7] showed that the number of characters not identified in the MJ character table is 52,294. These characters are classified into 3857 types (around 30,000 characters cannot be classified) and include 2037 types of wrong character. Thus the adoption of UTF-16 and IVS cannot solve the challenge of undefined characters used by municipalities.

Let us look at this issue from two perspectives which have been ignored in the past: cost and recognition. We would have to continue covering the cost of using an enormous number of characters even if we could resolve the Gaiji issue with the MJ character table, UTF-16, and IVS/IVD. Is it right that Japan's citizens should continue to cover this cost forever? Furthermore, are we even capable of recognizing and using more than 10,000 characters? This last question relates to the issue of human-machine interface.

4 The Cost Perspective

Some say that the issue of Kanji characters is spiritual and cultural and should not be approached from an economical perspective. But the number of people whose names include Gaiji is very small (about 4-5% of the population). The rest of the population must cover this cost through taxes or expensive products. We must consider whether this burden is justified or not. What is the cost?

I have calculated the cost based on surveys in three cities, the results of which are shown in Table 1. By economies of scale, big cities are more efficient than small cities, and the average city population in Japan is around 73,000. Therefore, the national economic loss is estimated to be between 1.2 billion yen and 2.7 billion yen, or around 2 billion yen.

The above economic loss is only a simple computation. There are many incalculable economic losses when it comes to municipalities:

[6] METI/IPA(2011),p74.
[7] Fuji Xerox Co., Ltd.(2012).

- Local governments must run high cost computer systems because they cannot change computer vendors easily due to Gaiji.
- Local governments must outsource even easy processing to computer vendors because standard PCs cannot handle Gaiji.

As a result, Japan must bear the burden of more than 2 billion yen in economic loss due solely to the issue of Gaiji in municipalities.

Table 1. Cost of Gaiji in municipalities

Cities		C City	F City	K City
Population		959,000	136,000	50,000
Detail (yen)	Support for claim	245,000	94,500	17,500
	Input existing Gaiji	3,733,333	126,000	175,000
	Register new Gaiji	1,085,000	63,000	87,500
	Maintaining system	710,643	1,000,000	14,000
	Data exchange with external system	2,450,000		26,250
	Data exchange with internal system	1,087,500		750,000
	Others	61,250	70,000	0
Total Cost (yen)		9,372,726	1,353,500	1,070,250
Estimated national cost(yen)		1,237,903,563	1,260,546,397	2,711,157,300

※The cost of one person-hour for a public officer is 3500 yen. The national cost is estimated proportionally from the above cities' populations.

5 The Recognition Perspective

5.1 The Recognition of Standard Kanji Characters

Using UTF16 and IVS/IVD might reduce these economic losses, but this must not be the main solution of this issue because it includes the problem of character recognition capability of human beings. In cases of man-to-man or machine-to-machine, the problem can be ignored, but in the case of man-to–machine, it becomes a big issue which cannot be ignored.

MEXT's Agency for Cultural Affairs published the Joyo Kanji table, "a guide for written kanji when used in everyday life in Japan." This table contains 2136 characters, and combined with the Hyogai Kanji table (characters not in general usage) gives about 3000 characters. Thus around 3000 characters are enough in everyday life.

Now let us look at Japanese Kanji recognition ability in the range of JIS level-1 and level-2 (about 6 thousand characters). Below are data showing how many people can understand and use (read and write) the characters of this range. The Japan Kanji

Proficiency Test Foundation operates Kanji literacy certification examinations three times a year. This association provides certification from 1st level to 10th level, 1st level being the range of JIS level-1 and level-2 (about 6000 characters) and pre-1st level being the range of JIS level-1 (about 3000 characters).

Figures 2 and 3 are the pass rates and the successful candidates of 1st and pre-1st levels from 2007 to 2012. The figures show that the pass rate of 1st and pre-1st levels is less than 0.25%, which means that even among people interested in Kanji, not even 0.25% can handle JIS level-1 Kanji (about 3000 characters). Successful candidates of 1st and pre-1st levels total about 18,000 people, or only 0.014% of Japanese, or about 1 person per 10,000.

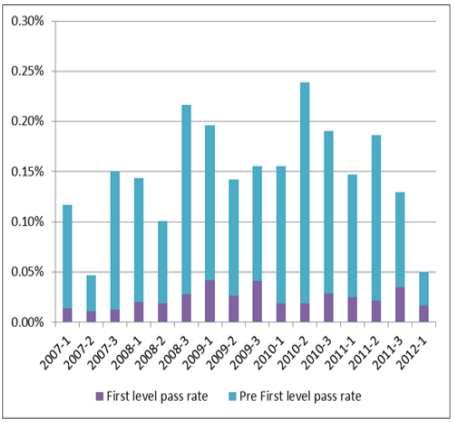

Fig. 2. Pass rate of 1st and pre-1st (left)

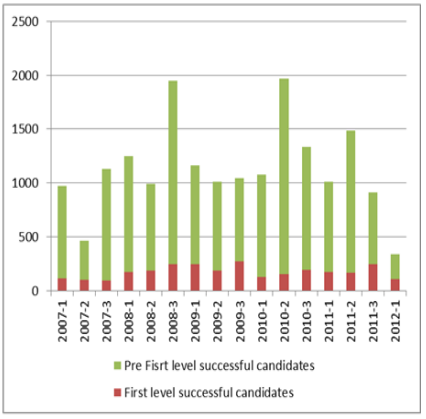

Fig. 3. Successful candidates of 1st and pre-1st (right)

Thus most Japanese people cannot understand and use even JIS level-1 (about 3000 characters) and only geniuses (2 persons per 100,000) can use JIS level-1 and level-2 (about 6000 characters). Why can Japanese use broader range of Kanji characters?

5.2 The Recognition of a Broad Range of Kanji Characters in the Field of Administrative Procedure

Next, to find out the Kanji recognition capabilities of Japanese people in the administrative range of Kanji (Juki and Koseki, about 60,000characters), I performed an experiment with check sheets. The check sheets were of two types: A used the range of JIS level-1 and B the range of Juki and Koseki. This experiment involved identifying pairs of Kanji characters arrayed in two lines of 12 in 2 minutes. The point of interest is the difference in recognition rates between A and B.

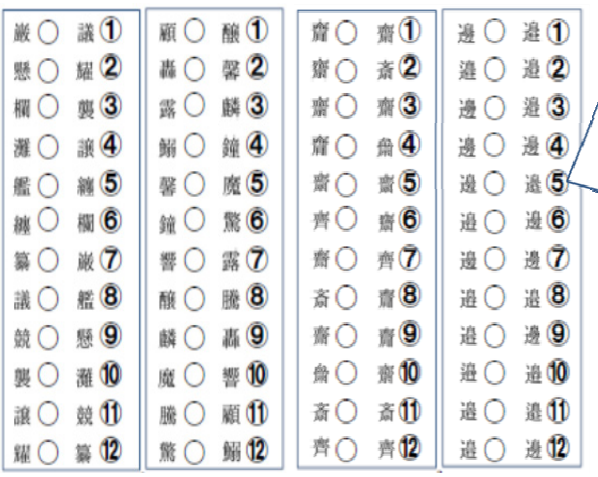

Fig. 4. sheet A (left)

Fig. 5. sheet B (right)

The examinees were of 3 types: private sector, local government, and general citizen. The level of education of each type was assumed to be higher than average. Elderly people were chosen to be general citizen examinees because the Japanese average length of life is 85 of female and 79 of male, and elderly examinees are likely to be active seniors who join lifelong learning courses. The results of the experiment are shown in Table 2.

Table 2. results of experiment on Japanese Kanji recognition

Group	Number of examinees (average age)	Sheet A (%)			Sheet B (%)		
		Right answer	Wrong answer	No answer	Right answer	Wrong answer	No answer
Local Government	96 (40's)	99	1	0	37	15	47
Private Sector	60 (50's)	98	0	2	33	27	40
General Citizen	47 (60's)	96	2	3	27	24	49

Table 2 shows the following:

- Most people recognized nearly 100% correctly on sheet A.
- People checked only one third correctly on sheet B.
- People checked 15-27% incorrectly on sheet B.
- People were unable to check 40-49% within the time limit on sheet B.

Elderly people (general citizens) checked half incorrectly on sheet B.

These results show that even those Japanese people who checked sheet A 100% correctly took much more time and made far more mistakes on sheet B.

6 Conclusion

The followings are points which were discussed above:

- UTF-16 and IVS/IVD cannot solve the issue of undefined characters (about 50,000 characters) used at municipalities.
- The entire population should not have to cover the economic loss (more than 2 billion yen) due to Gaiji for a small part of the population.
- People can hardly understand and use even a set of 6000 characters (2 persons per 100,000).
- People incorrectly recognize half of the characters from the set of 60,000 and need much more time to do so.

In the past, politicians and scholars said that the issue of Kanji should not be discussed from the perspectives of economy and efficiency because it is a spiritual and cultural issue. To the extent that Kanji is used between people, this might be true. But when Kanji is used between human and machine in a widely informatized society, a new problem occurs.

In the case of man-to–man interface, we accept the ambiguity of Kanji by negotiating with each other. But in the case of machine-to-machine, machines cannot accept ambiguity; even 1 bit of difference in the code causes a computer to treat characters as different Kanji. Even if we deal with similar Kanjis with IVS, humans must still decide which are variant characters and which are identical characters and register them to IVS using code. In the case of man-to-machine, we face a bigger problem: Our eyes cannot completely recognize the details in a character that a machine can. As shown by the above experiment, the recognition of 60,000 characters is beyond human ability. In this sense, we cannot communicate with machines.

Furthermore, it is important to note the fact that Japanese people cannot understand and use even JIS level-1 characters (about 3000 characters). The experiment above shows that Japanese people cannot use the sets of Juki characters (about 21,000 characters) and Koseki characters (about 56,000 characters) in everyday life, as seen in increased recognition time and a higher rate of incorrect recognition. Administrative procedures must always be carried out speedily and correctly, so we must limit the use of Kanji characters. Simply increasing characters and using IVS/IVD clearly will not solve the problem.

I propose the following two solutions:

- Solution 1

The use of Kanji characters for people's names and geographical name should be limited by law to the range of JIS level-1 and level-2 in the field of administrative procedure. The characters outside of JIS level-1 and level-2 which are used for

people's names and place names must be mandatorily changed to similar characters within that range. Characters that cannot be so changed must be expressed with Hiragana or Katakana. Any use of Kanji characters outside of this range for people's names or geographical names should be punished by law.

- Solution 2

This solution involves adopting Solution 1 with the exception of Koseki procedures. Only Koseki procedures would allow people to preserve their identity by using Gaiji, while Juki and other administrative procedures would be subject to the rules of Solution 1. To identify a person between Koseki and Juki, the same Juki code (or My Number) would be attached to both their Koseki and Juki entries.

As a point of clarification, I do not insist that the effort to use more characters with computer is wrong. It is necessary to use several million characters when performing cultural research, for instance, old documents or unknown characters. I do insist, however, that using several million characters in daily life is wrong. We must consider the public interest, instead of oversimplifying the issue as one of technology versus culture.

References

1. EGDI, United Nations E Government Survey (2012), http://unpan1.un.org/intradoc/groups/public/documents/un/unpan048065.pdf
2. Tatsumi, S., Enami, T., Yoshida, K.: A comparative study of e-Government in Japan and Korea. In: Proceedings of International Conference on Business Management 2012, pp. 159–169 (2012)
3. METI/IPA, Report on Character Integration Infrastructure Project (March 25, 2011)
4. Fuji Xerox Co. Ltd., Report on Gaiji used by Municipalities (contracted by Ministry of Internal Affairs and Communications) (March 2012)
5. Japan Kanji Proficiency Test Foundation (2012), http://www.kanken.or.jp/index.php