

# Computing in Japanese what are the frontiers now?

Jim Breen  
Monash University  
([jim.breen@infotech.monash.edu.au](mailto:jim.breen@infotech.monash.edu.au))

## Introduction

A workshop on Computational Japanese Studies provides an opportunity to take stock of what exactly is computing in Japanese, and how it differs, if at all, from language-processing activities in other languages. Where such differences exist, it is appropriate to consider whether they are still relevant today, when there has been a massive investment in such things as internationalized “single binary” software and unified character sets. Attention is also needed as to where computer technology can have useful impact in Japanese studies, and identifying where priorities should be placed.

## Computing in Japanese

Why do we even talk about computing in Japanese or Computational Japanese Studies? We do not talk about Computational Dutch Studies, or Computational Italian Studies. In this author’s view, the standout reason for this is the Japanese orthographical system. The mixed *kanji/kana* system, combined with other aspects of the orthography, leads to a number of issues which played a significant part in the introduction of information technology in Japan. Among these issues are:

- (a) encoding of *kanji* and *kana* in files
- (b) representation of text (display, print, etc.)
- (c) input of text (by humans)
- (d) segmentation of text into lexemes
- (e) canonicalization of accepted variants

All of these, and in particular the first three, had a major impact on the adoption of IT in Japan, and are considered by many to have led to a slower uptake of IT than in other countries with equivalent levels of industrialization. The importance of the issues listed above can be seen from the time it took to address them comprehensively. (As a test, consider the situation if Japan had adopted a totally romanized writing system in the early 20<sup>th</sup> century, as did nations such as Turkey and Malaysia. None of the above would have been considered issues in the introduction and use of computing.)

**Encoding:** It took over two decades from the introduction of computing in Japan before a national standard for *kana* and *kanji* coding was established, and many years elapsed before it was widely adopted.

**Representation:** Having two orders magnitude more characters to deal with certainly strained the capacities of end-user facilities. Interim low-storage and complexity solutions such as *hankaku katakana*, were used for many years (and persist today.) Solutions only arrived with technological developments such as VLSI, low-cost storage and non-impact printing.

**Text Input:** This was a major issue for many years, with complex single-*kanji* selection systems persisting until the late 1980s and beyond. Unger in his 1987 book “The Fifth Generation Fallacy”[3] asserted that the main goal of that project, launched in 1981, was to overcome the problem of Japanese text input through heavy use of AI techniques.

## **Current Situation**

All of the issues listed above were comprehensively addressed at the technological level during the 1980s and 1990s, and with the impact of internationalization many of the solutions have been embedded as standard elements in software. For example, virtually all major operating systems releases now have as installation options full support for Japanese input, display and printing (along with support for many other languages and scripts). Most high-level languages support non-alphanumeric text handling. Japanese text segmentation, which in the 1980s was regarded by many as an intractable problem, can now be performed effectively by several open-source and commercial systems.

Thus, the majority of orthography-related issues which tended to dominate the early stages of computing in Japanese have been adequately resolved. Computing in Japanese can validly be seen as on the same footing as computing in languages using alphabets, and the focus of “computational Japanese” is now largely on issues related to the language itself.

## **The Frontiers**

While the application of computer technology to Japanese studies is now in a similar position to other languages and cultures, and in areas such as NLP faces the same challenges, there are several topics which are worthy of particular attention. Addressing these topics should be made a priority in the application of computer technology. The following is proposed as a short-list of “frontier” topic which could well do with attention:

- (a) Dictionaries. Sue Atkins noted over a decade ago that computerization seemed to have limited impact on the user aspects of dictionaries, even if they were available on CDROM, and that “underneath these superficial modernizations lurks the same old dictionary”[2]. Japanese probably has the highest density of dictionaries of any language, and certainly large numbers are available electronically, but despite, or perhaps because of standards like EPWING/JIS X 4081, access to and presentation of dictionary content is still largely a replication of paper dictionary techniques. Atkins proposed a number of areas where computational resources should be exploited in a “new-age” dictionary, including extensive user customization, use of hypertext, etc. There is certainly scope for study in this area. In addition, there is a paucity of lexicons that are readily and freely available for research. Effort should be put into extending the free lexicons that are available, or seeking the freeing up of sources such as the EDR collection of lexicons, which are currently too expensive for many researchers.
- (b) Corpora. Japanese is not particularly well served in the area of available corpora, and is especially poor in the area of parallel bilingual and multi-lingual texts. While modest numbers of bitexts can be identified, they tend to be under commercial restrictions, and are generally unavailable for wide exploitation. The establishment of a comprehensive and representative Japanese corpus, and in particular the assembly of accurate bitexts, should be a priority.
- (c) Computer-Assisted Language Learning (CALL). For some reason CALL seems to be a “difficult” area for research. There are myriads of systems around, many of them commercial, yet few seem to get past being yet-another flashcard or vocabulary drilling tool. There has been little real research into the efficacy of such systems. One suspects that the problem lies in the gap between language education specialists and designers/developers of software. Given the popularity of Japanese study worldwide, (and indeed the popularity of English study in Japan), there is surely scope for proper research into where CALL has the greatest potential and which types of CALL tools are most effective.
- (d) Text Searching. With WWW search engines playing an important part in modern life, it is important that Japanese text is handled appropriately. Leading search companies such as Google and Yahoo apply a common framework for all languages, which at times does not completely cope with aspects of Japanese orthography, such as multiple written forms of words[2]. There is ample scope for more work in this

area.

- (e) Machine Translation. In many ways MT into or out of Japanese is in no different a situation as other languages. There is a small number of reasonable, but expensive commercial systems (e.g. Fujitsu's ATLAS), and a large number of inexpensive but poorly performing systems. Most R&D work seems to take place in commercial organizations, and hence IP issues preclude significant sharing of lexicons, etc. or even significant publication of methodologies. The area of statistical MT, which perhaps is seeing more non-commercial activity than traditional techniques, is hampered by the limitations in the availability of Japanese-Other parallel texts and freely available lexicons. This area of MT would greatly benefit by the expansion of readily available corpora and lexicons, as suggested above.

## References

1. B.T.S. Atkins, *Bilingual Dictionaries: Past, Present and Future*, Euralex'96, reprinted in *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins*, Euralex, 2002.
2. J.W. Breen, *WWW Search Engines and Japanese Text*, Sixth Symposium on Natural Language Processing 2005 (SNLP 2005), Chiang Rai, Thailand, December 2005
3. J.M. Unger, *The Fifth Generation Fallacy: Why Japan is Betting Its Future on Artificial Intelligence*, Oxford University Press, 1987