# Encoding principles of the Japanese philological texts

TOYOSHIMA, Masayuki [1] / mtoyo@Lit.hokudai.ac.jp

Fédération international d'Informations et Documentation, 1994

The main source of the examples cited in this paper is 玉塵抄 (GYOKUJIN-SHÔ, 55 volumes), written by 惟高妙安 (IKÔ,MYÔAN, 1480–1567). GYOKUJIN-SHÔ is a commentary on the character-prosodic dictionary of Chinese 韻府群玉 (INPU-GUNGYOKU,13c.), and is known to be one of the most bulky books produced in this era.

## 1 Japanese hybrid script system

### 1.1 overview of the Japanese script

The mediaeval Japanese language has 4 script systems, namely

1. pure *KANJI* (Chinese) script — usage limited (official documents)
2. *KATAKANA*(Japanese abridged [1] syllabic script) — academic and religious usage together with the *KANJI* system
3. *HIRAGANA* (Japanese syllabic *plain* script) — wide usage in mixture with the *KANJI* systems
4. *ROMA-JI* (Latin alphabet) – exclusive ussage in early Christian documents

In contemporary Japanese, these 4 systems are used freely in mixture with each other as in :

| この | FID | 国際 | ドキュメンテーション | 会議 | (the FID conference) |
|---|---|---|---|---|---|
| *2 (HIRAGANA)* | *4 (Latin)* | *1 (KANJI)* | *3 (KATAKANA)* | *(KANJI)* | |

### 1.2 concurrent script system as a translation

Since in the 9th centuries, the Japanese language has a method of writing Chinese sentences, simultaneously translating into Japanese by the aid of the *KATAKANA* system, written in small sizes at the side of each Chinese

[1] Department of Japanese linguistics, Faculty of Letters,Hokkaido university, Sapporo, Japan

[1] so called, because the characters have their origin in abridged style of *KANJI*s.

character, henceforth the *ad scriptum* method. This *ad scriptum* Sino-Japanese translation originated in the commentaries to the Chinese Buddhist documents, but soon developed into the native script system of the Japanese language itself, ie. writing down Japanese sentences as if they were Chinese, adding Japanese *KATAKANA* system *ad scriptum* as well.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ク | | ム | | | | シ | 2 | 1 | | て | には | | を |
| 内獨思 | 惟 | 、正 | 報 | 四 | 恩 | 、不 | 如 | 出 | 山林 | 、 | 臨 | 行 | 正 | 道 |
| | 1 | | 2 | | 1 | | | て | | | | 2 | | 1 |

(東大寺諷誦文稿 (A draft of ritual in TÔDAIJI-temple, ca.830))

Here the Chinese character (*KANJI*) strings look like as if they were Chinese (though broken), but it is intended to be read as a Japanese sentence, with the aid of the *ad scriptum* notations by Japanese *KATAKANA*. Because the rudimental word order of Japanese (SOV) is considerably different from that of Chinese (SVO), digits are provided to tell the correct order to assure the Japanese reading.

This *ad scriptum* method of writing down Japanese, disguised in Chinese, and furnishing it with full of notations, is quite common throughout the 9th to 19th centuries, and is still alive today, though the usage is limited.

# 2 threads in script

## 2.1 a script is read not always in one direction

The elements in the text are normally assumed to be read in the order of appearance.

This obviously does not hold in the Sino-Japanese *ad scriptum* translation, where the script is written in the Chinese word order, but the reading is done in the Japanese word order.

This is not a rare case unique to the SJ translations. Witnesses can be found in the vernacular tutorials of the Latin language, where the native word order is designated by digits together with glosses.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 5 | 4 | 6 | 8 | 9 | 10 | 11 | 7 |
| Syntaxis, | quæ | latine | constructio | vocatur, | est | recta | partium | orationis | inter se | compositio |
| a syntaxe | a qual | em latim | construcçaõ | se chama | he | direita | das partes | da oraçaõ | entre si | hũa composiçaõ |

(Joam Nunes Freyre, 1676, Margens da syntaxe com a construiçam portuguez) [2]

## 2.2 strings may be read more than once

Strings in written texts are read only once, if they are read at all. But there are cases that this simple assumption does not hold.

1. the ‘Read-Twice’ letters

Some Chinese words/letters signifying negation, modality or comparison, ‘未’, ‘當’, ‘猶’, etc. are translated into discontinuous morphemes in Sino-Japanese translation, namely

| Chinese | Japanese reading | |
|---|---|---|
| 未 | IMADA ... ZU | (not ... yet) |
| 當 | MASANI ... BESHI | (should ... do so) |
| 猶 | NAO ... GOTOSHI | (just like ...) |

These letters are simply *read twice*, [3] carrying two readings *ad scriptum* :

| | | | | |
|---|---|---|---|---|
| | 1 マサニ | 4 | 2 ナムチカ | 3 ミに |
| eg. | 當 | 在 | 爾 | 躬 |
| | 5 ベシ | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5(twice) | |
| → read in the order of | 當 | 爾 | 躬 | 在 | 當 | . |

[2] Maruyama(1993)

[3] In rare cases, a character may be read 3 times.

This “read-twice” system developed in the 10th centuries, and still is in use in the traditional Sino-Japanese translation (and is still taught at high schools).

2. optional readings

When several interpretations are possible, the candidates are written down together at the Chinese string concerned.

1–29.11[4] 高祖$\overset{テイ}{\underset{タイ}{帝}}$紀兩点ソ
(帝 reads either “テイ (TEI)” or “タイ (TAI)” ; both possible)

Here the string 「帝紀」 is expected to be read twice.

7–4.3 便宜留田..., 留田ハ, $\boxed{留田^{ヲ}_{ニ}}$ヤラ, 心ヲ不知ソ
(the phrase“留田” can be read two ways... I don’t know which is correct)

Here, the phrase “留-田” (literary *stay - field*) can be interpreted dually :

(a) with ヲ (WO, accusative) – keep the field
(b) with ニ (NI, locative) – stay *in* the field

The two readings written side by side signify that the reader has the choice : the author himself admits he does not know which is correct.

He could have written :

1. 留田ヲヤラ, 2. 留田ニヤラ, ...

Instead, he puts the candidates on either side of the Kanji string,

留田$^{1.\ ヲ}_{2.\ ニ}$ヤラ, ...

and forces the reader to read the string(留田) twice, 1st time [5] accusative, 2nd time locative, giving both of the candidates.

This method of giving candidates concurrently is quite common in mediaeval Kanji dictionaries, where differences in usage or subtleties are indicated simply by their positions, sizes, or even colors (red and black) of the candidates.

3. intentional overlapping (“bridge-words”)

A word (morpheme) may represent the ending of a sentence and the beginning of a new one at the same time. This pun is especially loved in the traditional verse.

TARE-NI YUKUYE -WO *TÔ*TÔMI (誰に行方を遠江)
*whom destination acc. ask*
(I don’t know whom to ask (*TÔ*) the destination, and am going to that country named *TÔTÔMI*...) [6]

[4] cited from the main source 玉塵抄 (GYOKUJIN-SHÔ).

[5] The order is not relevant; maybe indeterminate.

[6] Paul Claudel(1929) tries to translate this very phrase into French with the same effect as
*Notre chemin à qui le deman-/ des Estuaires lointains...*
where ‘*(deman)der*’ and ‘*des*’ are intentionally overlapped.

These overlapping morphems are not literally read twice, but their grammatical analysis calls for one by one treatment.

# 3 SGML/TEI assumptions on descriptive markup revisited

## 3.1 descriptive markup is subjective

The principle of descriptive markup basically assumes that the appearance of a document and their effects derive from its logical structure of the elements. ie. the content determines its appearance, the appearance being a mere result.

This assumption leads to an inherently subjective markup, because the encoder must perceive from the appearance, the intent of the author who selected that very rendition, especially when one single notation is used to express several types of contents.

According to the “Hart’s rule” (Oxford UP,1978) an author may use *italicized* type-faces, in order to express titles of books, pictures, music, names of ships (but not dogs), stage directions, and simply to emphasize. So the encoder is expected to be able to tell whether the name ‘*Titanic*’ is printed italic because it is a ship name (ie. no intension of emphasis), or it is an emphatic call to a (huge) dog and thus must be encoded with `<emph>`.

The descriptive markup assumes that the encoder understands why the content takes that specific appearance, and expects the encoder to link both by giving them a suitable tag representing elements / attributes happily. This is typically achieved when the author is also the encoder, where the subjective markup is not a problem, but welcome.

## 3.2 encoding as an interpretation

This basic assumption of descriptive markup sometimes fails to hold, not because of the insufficient or inadequate comprehension on the part of the encoder, but because the author himself is not sure what to write.

This happens when the text is a commentary on some other document, and the author-commentator cannot understand the original.

In GYOKU-JIN-SHO(玉塵抄), the author IKO(惟高) frankly admits that he cannot understand the base document INPU-GUNGYOKU(韻府群玉).

1–97.7「涙可汗金銀」ノ五字エ心エヌソ、ナニト点シテ心エウスヤラソ
(「...」these 5 letters are hard to understand. Do not know how to read them.)

7–197.1「善肉声羅蛟挑之不答⌃」、此事心エス、点モ$\overset{クダシ}{下}$カタイソ、除之
(「...」, Cannot understand this. Do not even know where to begin a sentence and where to end it. Let’s skip this.)

This happens sometimes by a distorted original,

7–226.6 鐘□ノ下マメツシテミエヌソ, 名ヲシラヌソ,
(cannot read the letter below「鐘」, it’s worn out. )

7–162.12 此句モ$\boxed{含}$ノ字マメツシテ正字ヲシラヌホドニヨメヌソ、...$\boxed{含}$ノ字含カ合カミエヌソ
(the letter is worn out and cannot tell whether「含」or「合」)

or by the pure lack of accessibility to the text referred.

1–36.9 桑東... イワレアリサウナゾ, 此モ北史ノ語ゾ, 北史ナイホドニ, エカンガヘヌゾ (This “桑東” surely refers to some anecdote in BeiShi, but this book is not around, and I have no idea.)

7–165.14「便宜」モ$\overset{タヨリ(ヨロシキ)}{便\quad宜}$ニト云心カ, 漢書ナイホドニ不考ナリ
(「便宜」may read differently according to the context..., but don’t have the original HanShu, thus am not certain.)

One may argue here, that the descriptive markup is useless, if not impossible, because it is clear that there is no intent whatsoever on the part of the author.

But the descriptive markup need not always be a faithful representation of the intent of the author. The markup represents the interpretation of the encoder, not the author.

## 3.3 model of elements and attributes

The SGML language, which lacks semantics [7] , is based on the assumption that the structures and their values can be, and should be divided. This is a direct offspring of the *data abstraction*, where the data (instances) shall be handled only through previously well-defined manipulations, and the direct values (*LITERAL*s) should be deliberately hided from those manipulations.

According to this view, what counts in the meaning is the structure (`ELEMENT` in SGML) not the value (`ATTRIBUTE` in SGML). The value is given to the terminal (the lowest) elements in the hierarchy, but the hierarchical meaning of the structure is expressed in the element hierarchy itself : ie. meaning of an element is determined by its position in the structure, not by its *LITERAL*. The SGML language does permit `ELEMENT`s to be composed of lower `ELEMENT`s, and one can define `element`s of `element`s, or `element`s of `attribute`s, but not `attribute`s of `element`s. nor `attribute`s of `attribute`s.

This assumption encounters a problem, where the values(LITERAL) themselves have structures.

1. proper nouns

The simplest example is a certain sequence in the human names.

- One of the common Japanese male names is a composition of a digit and 「郎」(man / son). eg. 「一郎」(1st - son),「二郎」(2nd - son),「三郎」(3rd - son), etc., thus we can discern from a name, his order of birth.
- A specific character is conferred as an honorable title.
eg. titles of the great performers of the KANZE(観世) NÔ theater : 「紅雪」(red - snow),「華雪」(brilliant - snow),「雅雪」(elegant - snow), etc. If a performer has '雪'(snow) in his name, he is honored.
- Or a tradition is shown by a succession of a specific character in names.
eg. the names of the tycoons of EDO (17c. – 19c.) period,「家康」(1st),「家光」(3rd),「家綱」(4th),「家宣」(6th),etc.

2. notes on notes

Notes can be given to a Japanese reading of a Chinese character.

ス ム
タン
1–10.11 潭 ト云フ小国アリ,

(there is a small country called 「潭」)

Here, first the word「潭」is given a Japanese reading「タン」(TAN), and then a note is added to the reading「スム」(unvoiced), ie. it's not DAN, but TAN. Clearly the note (which is itself an element) is not given to the word (element), but to the value (attribute).

テイ　　　ス ム ス ム　　　　　　　　　　ヂヤウゲン
1–8.6鄭玄モ本テハ, テイケント, ヨムソ, ソラニ云時ハ, 鄭 玄トヨムソ

(in books, the name 「鄭玄」reads unvoiced テイケン (TEI-KEN), but in speech, reads voiced as ヂヤウゲン (DIÔ-GEN))

[7] TEI is one of the set of agreements on its semantics.

Prosody in verse may arise the same kind of problems : eg. Rhymes (especially the so called *eye*-rhymes) are discernible only through their attributes (ie. the written/pronounced form). [8]

## 3.4 problems peculiar to SGML

As TEI relies heavily on the SGML language, it takes over the inconveniences of the SGML as well.

1. No block

SGML lacks syntactic blocks. As a result, an SGML document cannot (readily) include another SGML document. This necessitated the `SUBDOC` feature, which is not a block at all, but a completely independent document, sharing no information with the caller.

This is a serious problem in composing a large SGML document from smaller ones, especially in constructing corpora.

2. No scope

Because of block-less-ness, the name space of an SGML document is flat. ie. names should not duplicate in a document, except for the attributes.

This means that when one starts to encode a document fragment, and if he tries to setup `ELEMENT`s of his own, he must make a painstaking effort to scrutinize all the DTDs and the `ENTITY` definitions he plans to include, in order to avoid name collision.

Contemporary programming languages, especially OOP (object-oriented paradigm) ones, have overcome this problem by scope rules and inheritance, but SGML has neither.

## 3.5 attribute inheritance

As an SGML document is a flat plane, there is no hierarchy unless explicitly so stated.

Even if it is so stated, the descendents of an upper hierarchy do *not* inherit the attributes of their ancestor, because attributes are property of a specific node of the hierarchy tree, not that whole branch.

For example, if one wants to setup an `ELEMENT` representing *damage*s, and wants to divide the element into sub-classes of damages caused by water, worms, burnt, etc., the *water-damage* element knows nothing of the simple *damage* element, its mother.

This eventually bars the hierarchical encoding of structures, and leads to a bunch of mutually overlapping `ELEMENT` and `ATTLIST` definitions instead. [9]

TEI tries to solve this problem by defining variety of attributes instead of defining them as proper elements, by including the same attributes (`%a.global`, etc.) at every level of a tree, and by simply (ie. manually) replicating the upper attribute into lower [10] to share information in different nodes of a branch.

Aside from the danger of making unnecessarily global the scope of attributes (which is practically the sole secure name-scope in SGML), this method inhibits the attribute overriding (ie. peculiar specifications and reformulations at levels), which is a *must* item of an OOP language.

### 3.5.1 validation

[8] TEI-P3 provides a method of pointing and mutually linking rhyming elements. (9.5/p.265).

[9] TEI-P3 defines over 400 `ELEMENT`s.

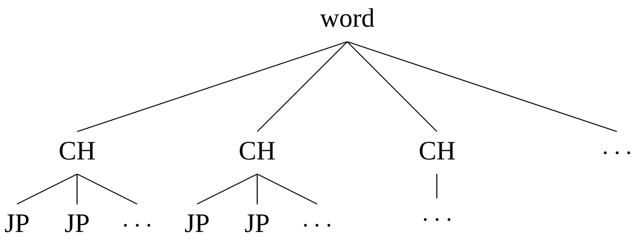
[10] TEI-P3 3.7.1/p.59.

SGML does not provide any facility for the validation of `attributes` (NB. it does for the `elements`). As TEI-P3 makes an extensive use of attributes, not-applicable, contradicting, or nonsense attributes, as well as dangling pointers, should be flagged, not by SGML parsers, but by attribute validators tuned for the TEI-P3, which are yet to appear.

## 3.6 elements of different hierarchies

The SGML/TEI model expects that the abstract element structure (hierarchy) is fairly stable, that one DTD (document type definition) can be used in a set of documents. Thus, according to TEI-P3, the `<front>` (preface, dedication, etc.) always comes before the `<body>` of a novel, and the `<back>` (if any) always follows it. [11]

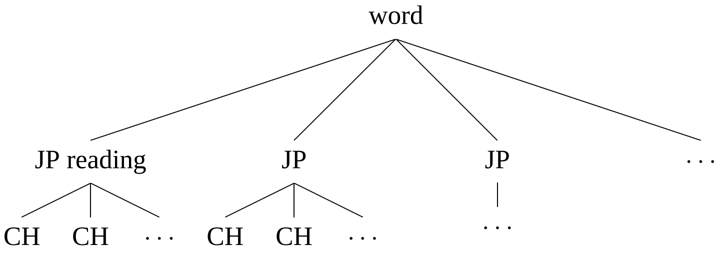
In Sino-Japanese script, a word is represented by several Chinese characters, to which Japanese readings are attached.



But it is not uncommon that this hierarchy is overturned, Japanese strings being notated by Chinese characters, called *attaching Kanji*s.

平
1–75.2ヘイサブライヲ, ヨブニハ, 弓ヲ以テスルソ

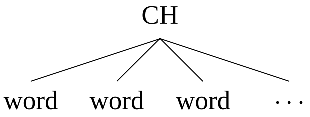
to clarify the meaning of “ヘイ”(HEI), a Chinese character “平” (common, plain) is attached.



or even the words are employed to the notation of a character

シン
9–461.5榛ノ木デスルソ
ハシハミト云ソ

the reading of “榛” is given by the attaching KANA “シン”(SHIN), and another reading in native Japanese is given by a sentence, “ハシハミト云ソ” (commonly named HASHIBAMI) employing Chinese characters .



[11] *Tristram Shandy* is an apparent counter-example, where the *preface* comes between the chapter 21 and 22 of the volume 3.

These structures can be comprised into DTDs, as SGML does permit a recursive structure, where an element refers to itself directly or indirectly. But defining elements recursively only to achieve the reverse hierarchy, is rather an abuse, because none of the examples above is truly recursive ; they are just upside-down.

# 4 problems in the distribution of e-texts

## 4.1 problem of coded character sets

The most widely used coded character sets for the contemporary Japanese are those by the Japan Industrial Standard (JIS)

1. `X0208-1990` Code of the Japanese graphic character set for information interchange
2. `X0212-1990` Code of the supplementary Japanese graphic character set for information interchange

both of which the revisions are planned to be published in 1995.

Quite unfortunately, X0208 went through a revision in 1983, where no less than 250 characters are replaced without compatibility with the forgoing versions, although the revision board (at that time) insists in claiming that they touched only the appearances of the characters. [12] This means that the e-texts compiled *before* the revision may be read differently in current standard. This really happens when printing devices made *after* the revision are employed, and as a result, current printers are equipped with `old-JIS/new-JIS` switches. [13]

The current revision board, of which the author of this paper is a member, pays special attention to remedy this confusion, by providing a full description and explanation for each code-point, together with pointers to other KANJI dictionaries, and (possibly) the permissible variations of its representation (glyphs), of which information the forgoing standards lacked nearly completely.

Currently, manipulation of Japanese e-texts (especially those compiled before 1983) will be a mess without an appropriate note that mentions in which version of JIS it is encoded. The `Writing-system declaration` of TEI-P3 is a monumental achievement for the improvement of this (lamentable) status.

The JIS standards offers (only) 12,000 KANJI characters, which proves to be too scarce for the encoding of Japanese classics, especially those Sino-Japanese scripts. Though the UCS(Universal character set) of `ISO/IEC 10646-1` Universal multiple-octet coded character set (BMP – basic multilingual plane) came out with over 20,000 KANJI's, the philological community in Japan is not satisfied and still wants to add tens of thousands of KANJIs to the other planes of `ISO 10646`.

This must not be done without giving an explicit principle in identification and discrimination of KANJIs, which is a hard task yet to be solved in the Japanese philology.

## 4.2 freedom of contribution and distribution

The Japanese script of classics, especially hand-written ones, are no less hard to read to the human eyes than to the OCR(optical character reader)s. This explains the apparent scarcity of the Japanese e-texts.

This is a blessing in disguise however, because the e-texts have to be manually encoded only by the specialists in that field (because they are the ones who really need them), thus prove to be genuine in quality, and completely free of copyrights (except for that of the encoder).

The cost of communication have long barred the distribution of the encoded texts of Japanese classics, but the advent of the Internet has greatly improved this situation recently. Each researcher will soon be equipped with the full ability to distribute his/her own versions of a classics.

[12] cf. Nomura(1984)

[13] cf. Toyoshima(1992)

This means that the contribution of encoded texts to the community, is no longer a privilege of select scholars, but has become the right of everyone who has interest in it. Anyone in the community can freely make use of the contributed property without impairing or infringing the right of the author / encoder, and everyone shall be entitled to further distribute the gift, in addition. In other words, the freedom of contribution and distribution is the necessary condition for the e-text encoding of classics.

## 5 references

Claudel, Paul(1929) L'oiseau noir dans le soleil levant (Gallimard)

Maruyama, Tôru(1993) Christian documents as a linguistic monument in the great navigators' era［in Japanese］(*Nanzan-Kokubun-Ronshu*, 17, Nanzan university, Nagoya-shi, Japan)

Nomura,Masa-aki(1984) The revision of `JIS C6226` Kanji character set.［in Japanese］(*Hyojunka journal*, 14-3, Institute for the standardisation of Japan)

Oxford UP(1978) Hart's rules for compositors and readers at the university press Oxford (Oxford UP)

TEI-P3(1994) Guidelines for electronic text encoding and interchange edited by C.M.Sperberg-McQueen & L. Burnard, (ACH/ACL/ALLC), circulated by Text Encoding Initiative, also available by anonymous FTPs at `sgml1.ex.ac.uk`, `ftp.ifi.uio.no`, and other TEI-distribution sites. ),［cited by the chapter.section number and the pages of the printed version(total 1290 pages) on Apr.8, 1994］

Toyoshima,Masayuki (1992) On the so-called "characters not found in JIS "［in Japanese］(*Sinica* 3-2, Taishukan-shoten)

## 6 acknowledgements

I would like to express my sincere gratitude for professor C.M.Sperberg-McQueen and his assistant, for having sent me the TEI-P3 draft for consultation before its publication.

I would like to thank professor Syun Tsuchiya, and professor Mari Nagase, for having offered me a chance to be acquainted with the TEI project.

I acknowledge that this paper is based on the grand-in-aid research (`#C-03610221`) funded by the Ministry of Education, conducted by professor Asako Izumo, to whom very much grateful I remain.

(EOF / M.Toyoshima)