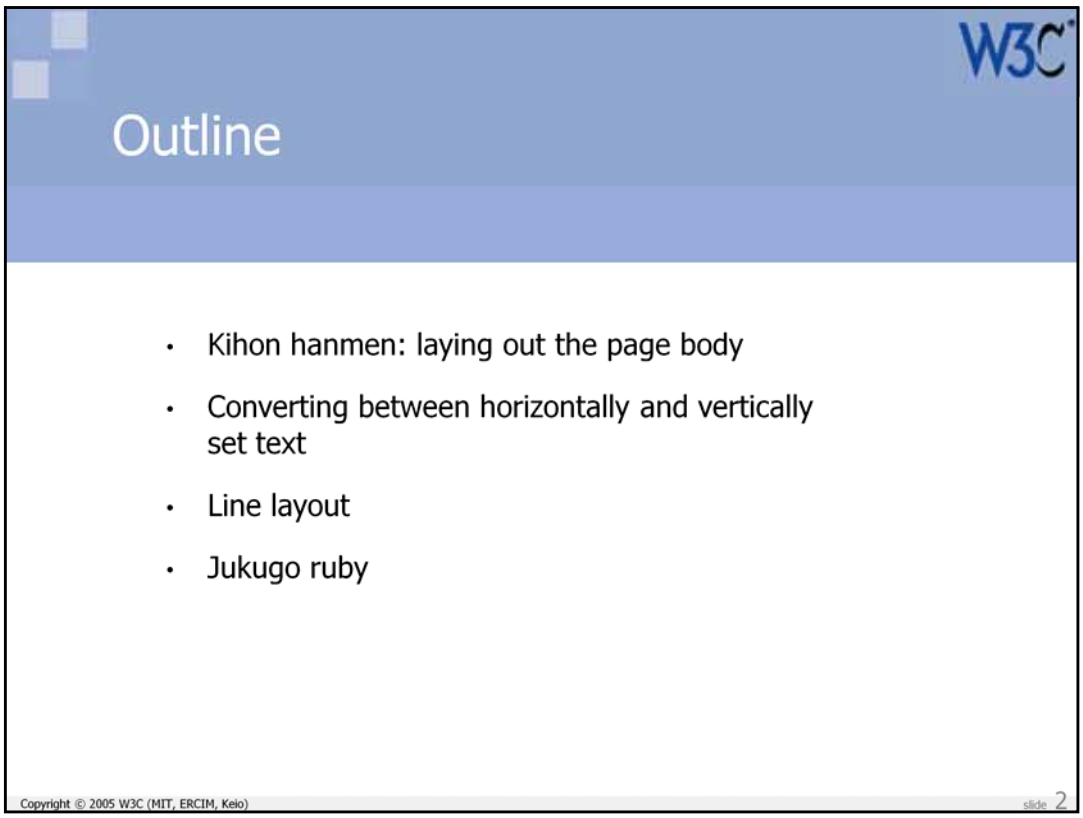
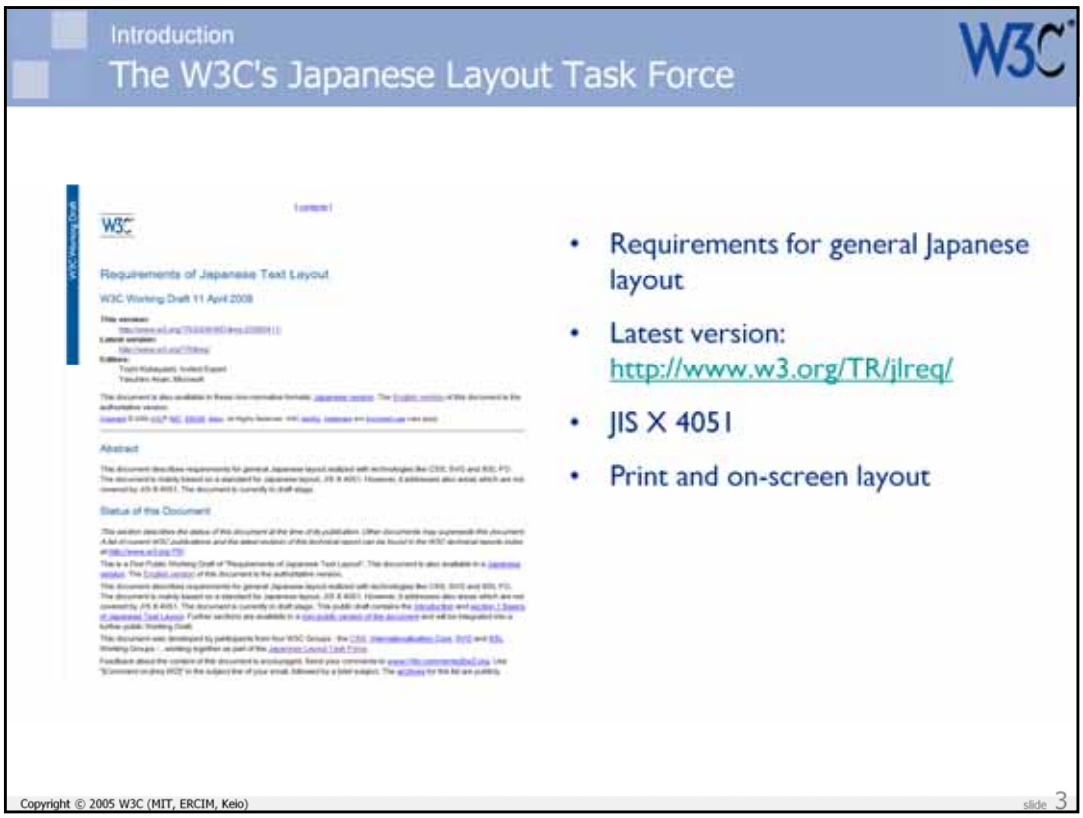
W
3
C
New Work on Japanese Layout Requirements
Richard Ishida, W3C
Tatsuo Kobayashi, JustSystem
Steve Zilles, Adobe Systems


Outline
Kihon hanmen: laying out the page body
Converting between horizontally and vertically set text
Line layout
Jukugo ruby
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 2



The goal of this task force is the creation of a document about requirements for general Japanese layout realized with technologies like CSS, SVG and XSL-FO. The document will be mainly based on a standard for Japanese layout, JIS X 4051. However, it will address also areas which are not covered by JIS X 4051. The document is currently in draft stage and is being developed further by the Japanese participants in the task force.

The task force will gather and integrate feedback from the participating working groups about the need and technical feasability of various requirements.

The output of the task force will be an English document (a Working Group Note) describing requirements for Japanese Layout. The description will not be on a technology-specific level. It is assumed that the participating working groups (see below) will decide independently how (and if) to implement the requirements in their specifications.

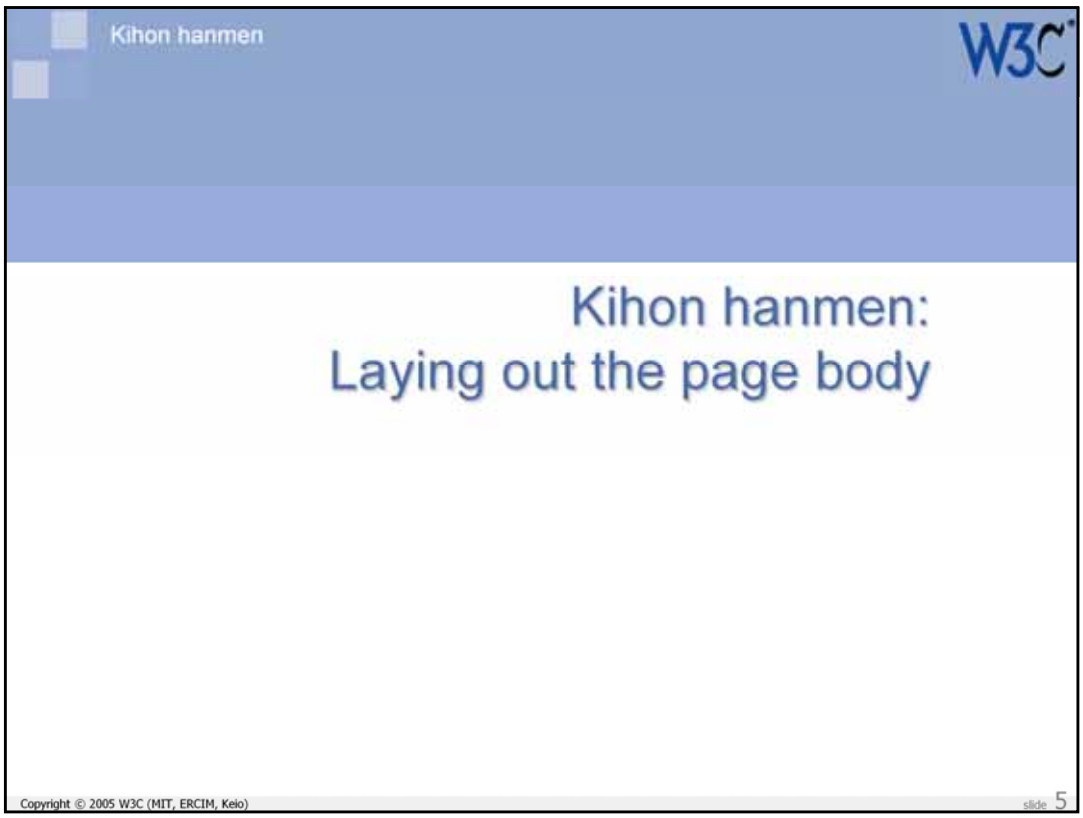
Introduction

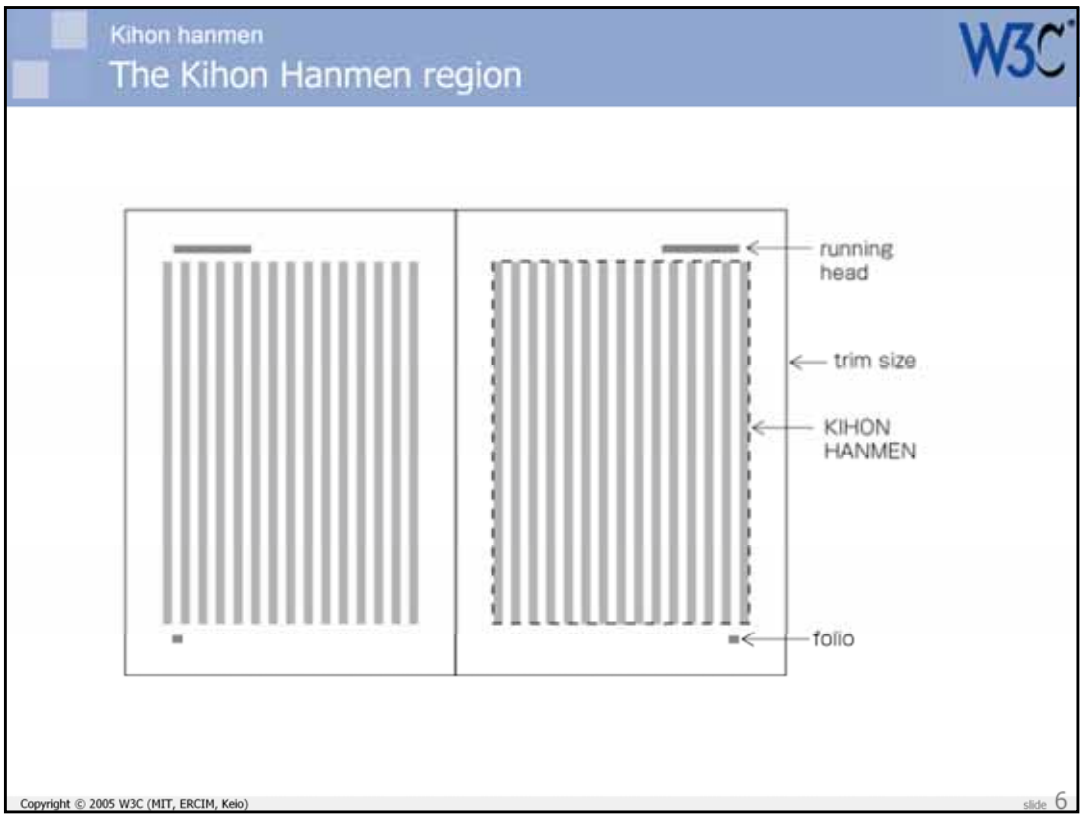
The W3C's Japanese Layout Task Force

- Editors: Toshi Kobayashi, JustSystems, Yasuhiro Anan, Microsoft
- Participants from I18n, CSS, SVG, XSL Working Groups
- Working in Japanese, with periodic meetings in English
- Output: W3C Note (in English and Japanese)

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 4

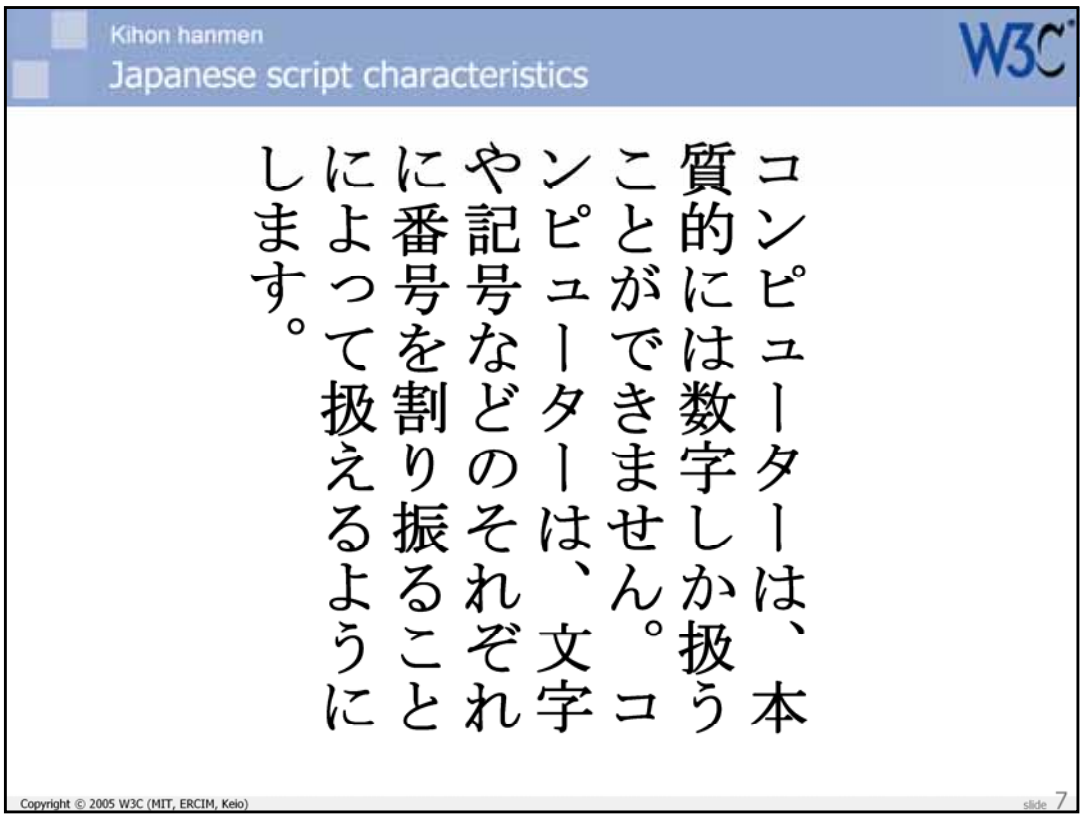

Kihon hanmen
Kihon hanmen:
Laying out the page body
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 5



The Kihon hanmen is the area of the page that contains the main body of the Japanese text. It doesn't include page headers and footers.

The document in the slide is vertically set and typical of the layout of a Japanese novel.

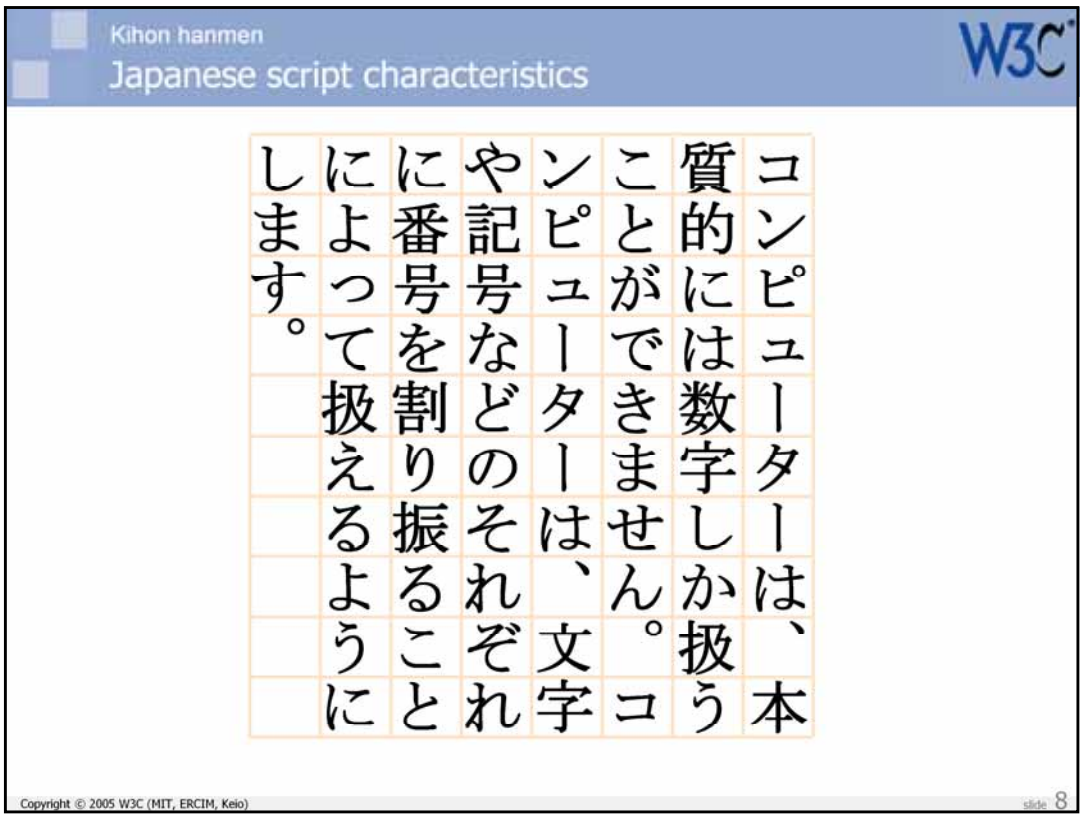
In order to understand why the kihon hanmen is of interest, we must first understand some of the characteristics of Japanese characters and their use.



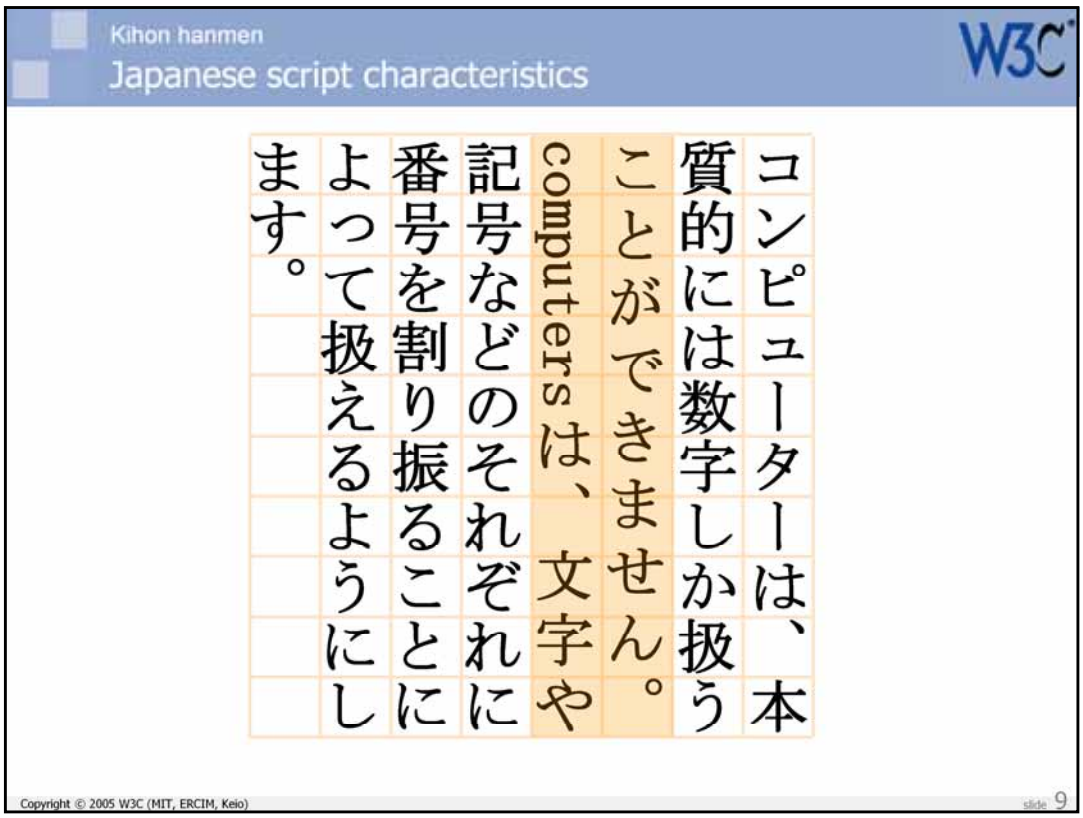
Japanese kanji, hiragana and katakana characters normally sit in small square boxes that are all the same size. This leads to a very regular effect.

Even characters such as the Japanese period and comma shown on the slide sit within a square box of the same size.

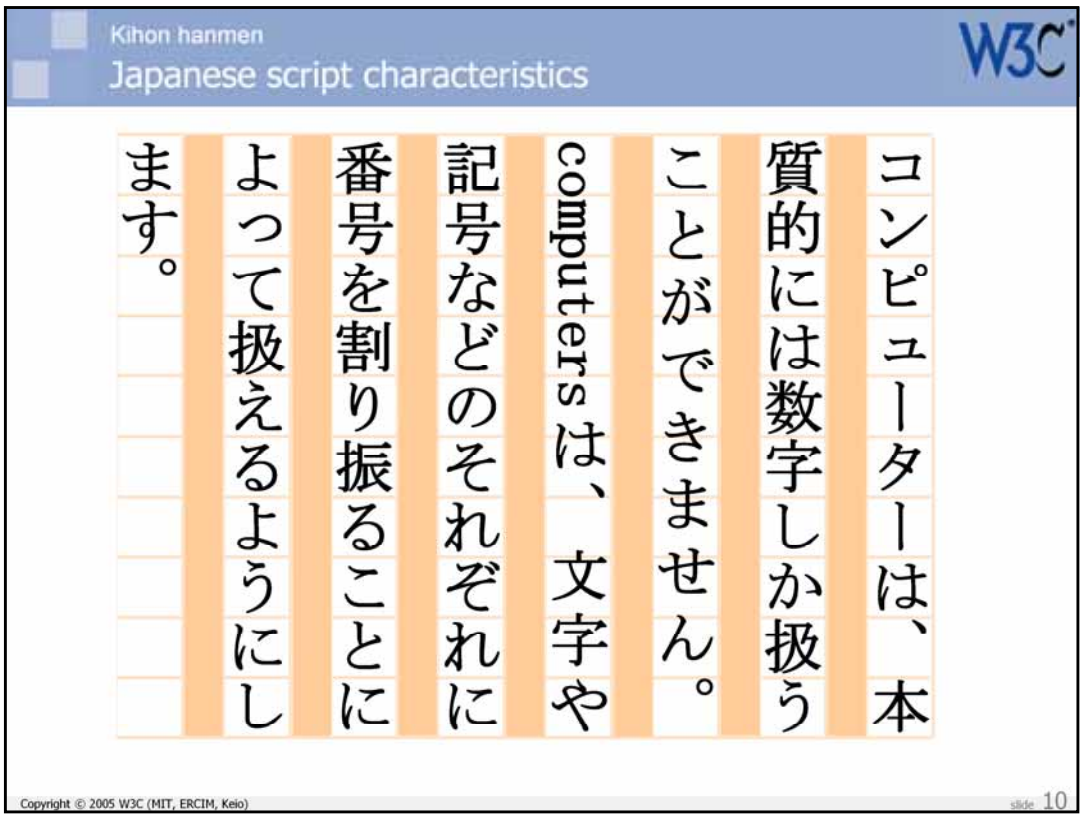
There are no spaces between characters in the slide above.



The regularity of the character sizes produces a grid-like effect. Lines could be drawn through the text both vertically and horizontally, as shown.

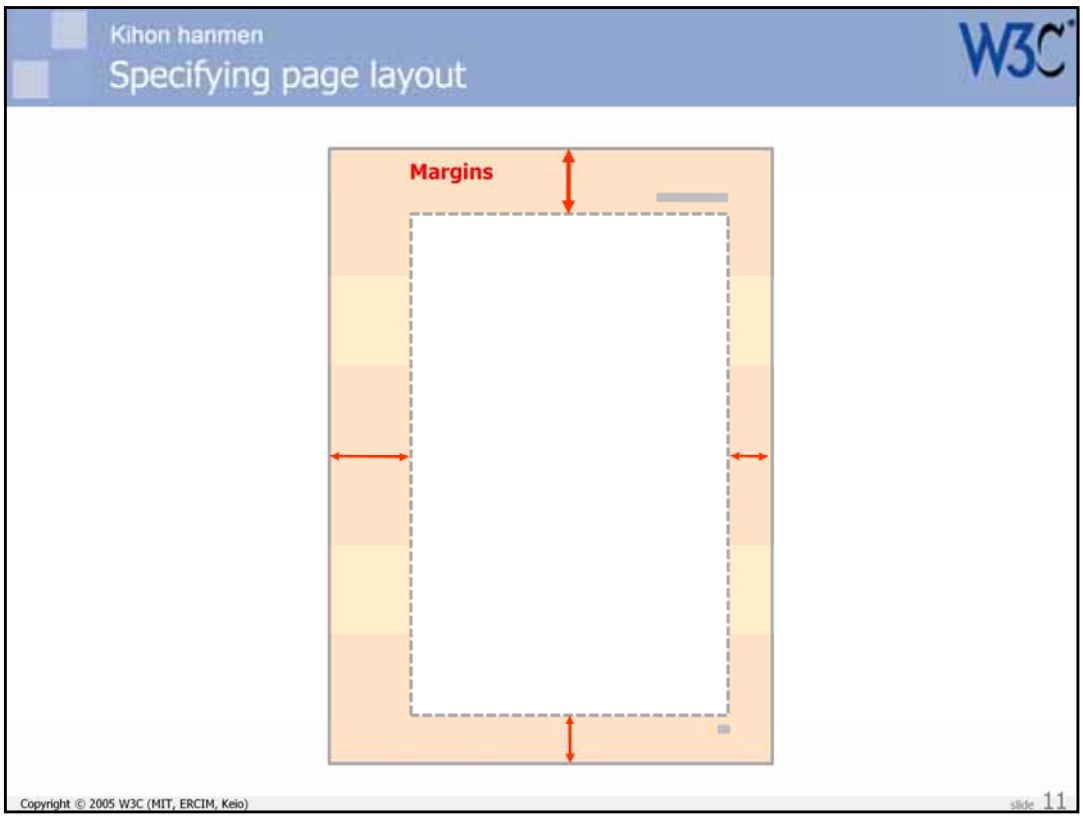


In reality, life is not quite so simple. Embedded Latin characters, rules about what characters can begin or end a line, etc, tend to produce small variations in the placement of characters relative to the grid. Nevertheless, where characters come adrift from the grid cells, there is a strong tendency to return to the grid pattern as soon as possible. This is achieved using justification rules. If the grid layout is not recovered before, the last character box on a line will end flush with the bottom of the grid.



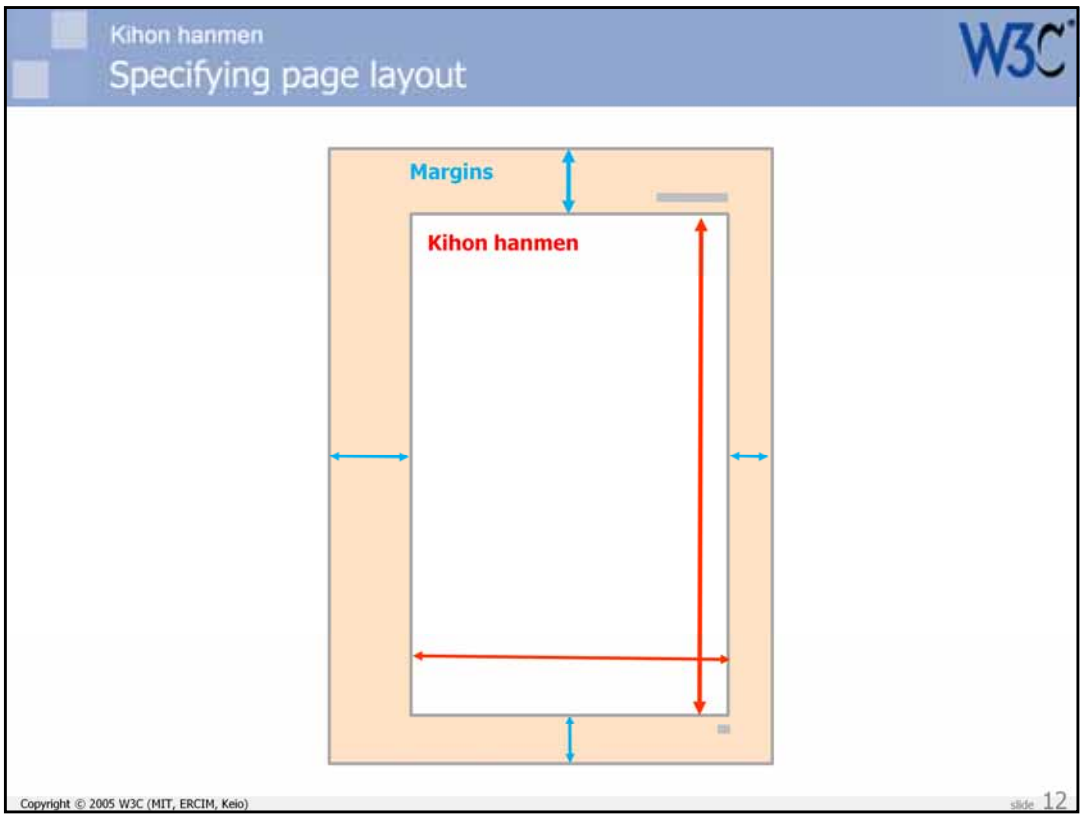
Actually, the grid cells are not usually square. Japanese text usually has some interline spacing, typically ranging from half a character box in width, for short lines, to a full character box for longer lines.

This tendency for characters on a page to fit to a grid makes the kihon hanmen a little special in a couple of ways, as we will now see.



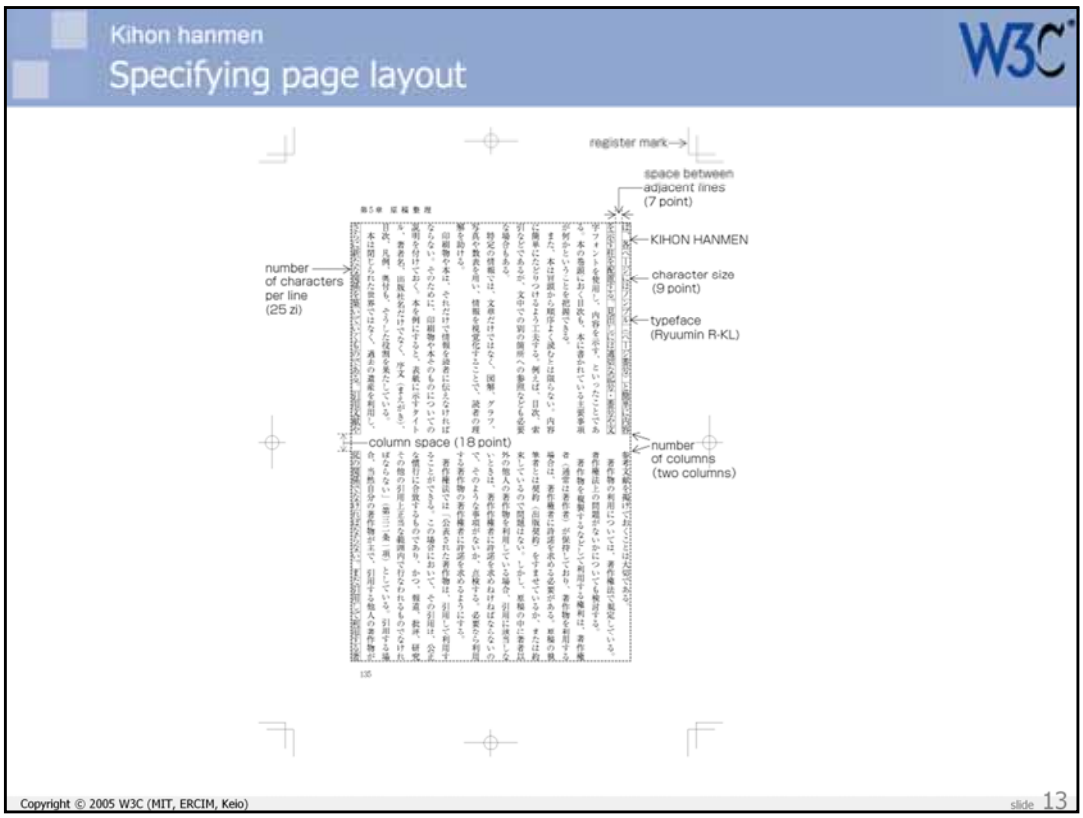
A typically Western way of creating the main text region on a page involves just specification of margins. For example, to create a simple document in XSL-FO, after specifying the printable area of the page using page margins, you specify the size, within the printable area, of the main body region by specifying its margins. (Those margins can then be used for page headers and footers.)

The key point is that you don't usually specify the size of the body region itself. The size of the body region is whatever is left over after you have specified its margins. In other words, you specify the size of the blank space, then just flow the text into the remaining space.



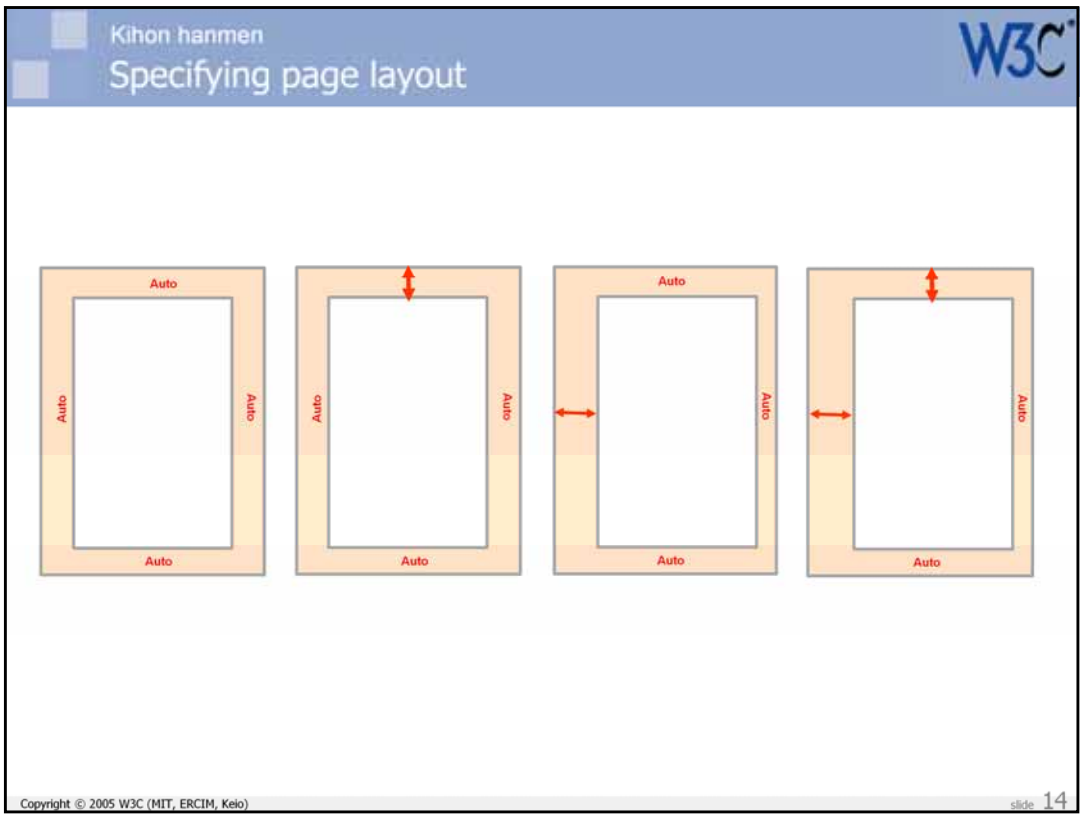
Because the kihon hanmen is a grid, Japanese people think of page layout slightly differently.

Their natural inclination is to first define the size of the main text area (ie. the kihon hanmen). Only after that are the margins considered.



For a page with a single column, defining the size of the kihon hanmen involves choosing a character size, specifying the length of the lines as a number of characters, then specifying the other dimension of the text area as the number of lines per page and the size of the interline gap. (There are preferred sizes and ratios for characters, line lengths and lines per page.)

For documents with multiple columns, the size of the gap between the columns is also taken into account.



Strategies for determining margins.

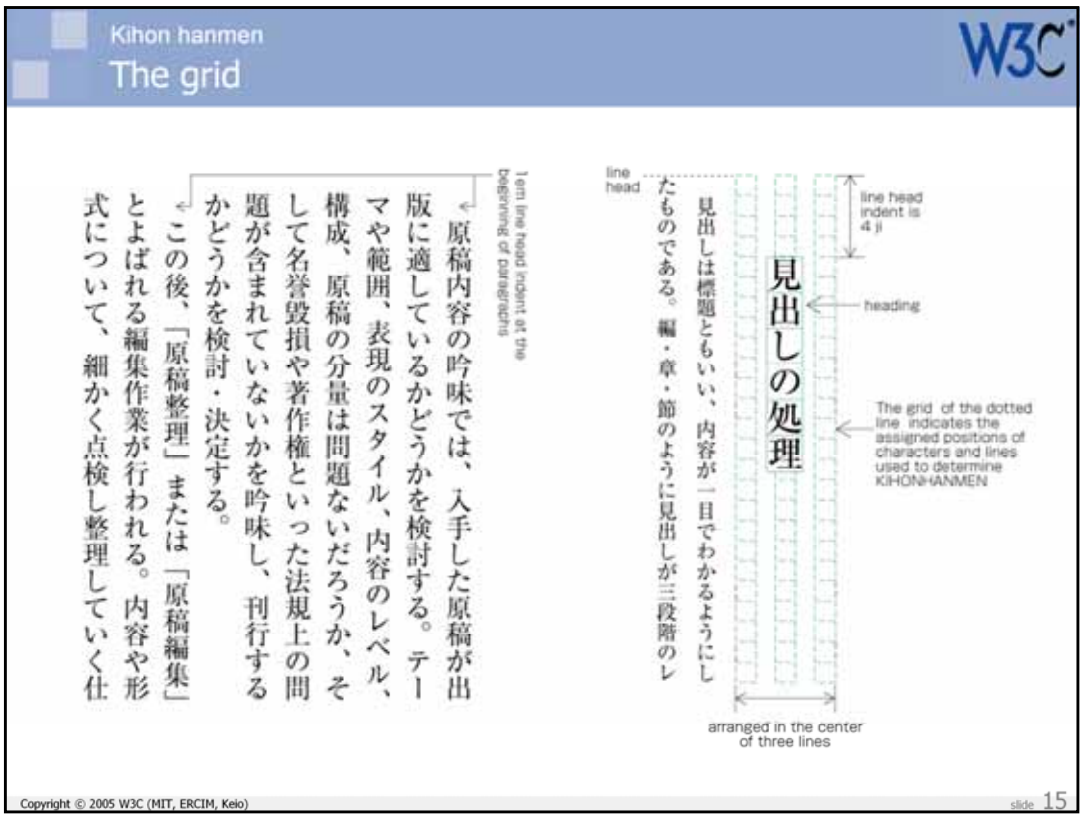
The common default is to center the kihon hanmen in the body region.

Other alternatives include:

1. Specifying the top space (for horizontally set pages) or the bottom (for vertically set pages), and centering the sides.
2. Specifying the size of the gutter to the side, and centering the top and bottom.
3. A mixture of 1 and 2.

In summary, to specify the position of the main text flow in a simple document using kihon hanmen you would specify the dimensions of the main text region itself, and zero to two margins.

There has been discussion at the W3C about making it easier to specify the body region in this way in XSL-FO 2.0.



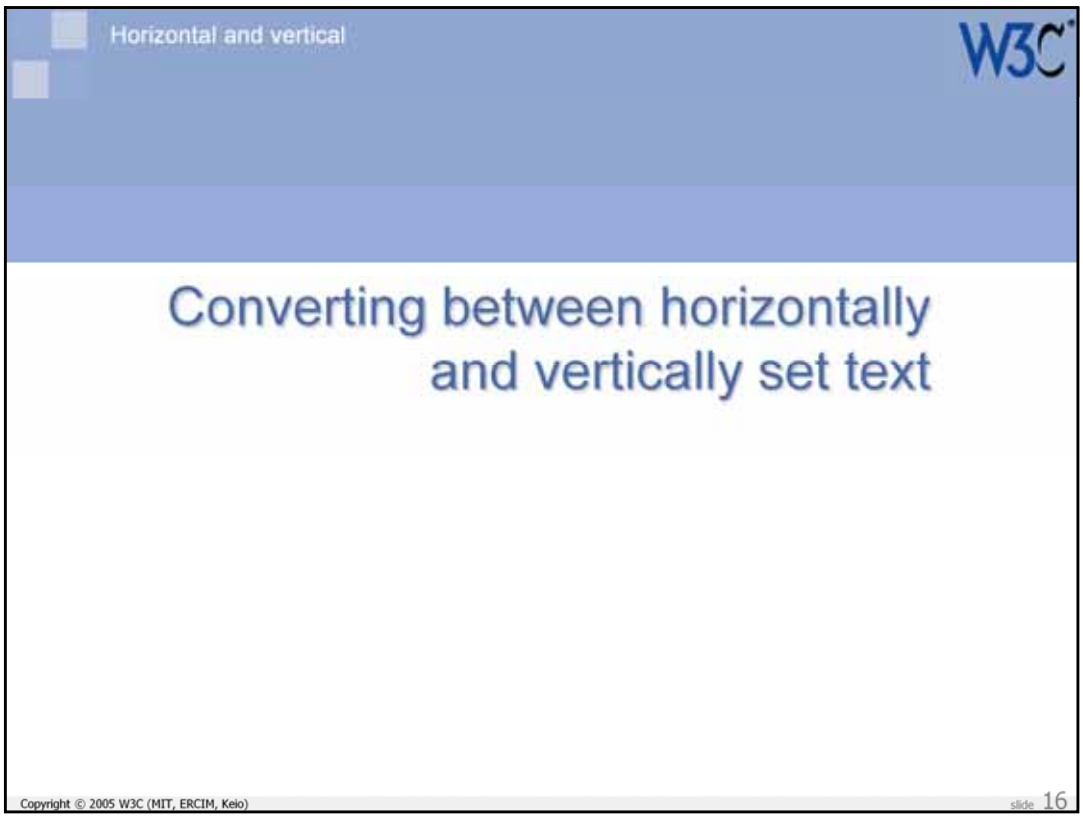
Another significant characteristic of the kihon hanmen as a type of body region goes back to it's grid-like nature.

Elements positioned within the kihon hanmen tend to snap to the logical grid that underlies this region.

For example, the first line of a paragraphs tends to be indented, and the usual indentation is one character box.

Headings in the document also tend to be positioned relative to the grid, even though the size of the characters in the title is larger than that used for the logical grid of the kihon hanmen.

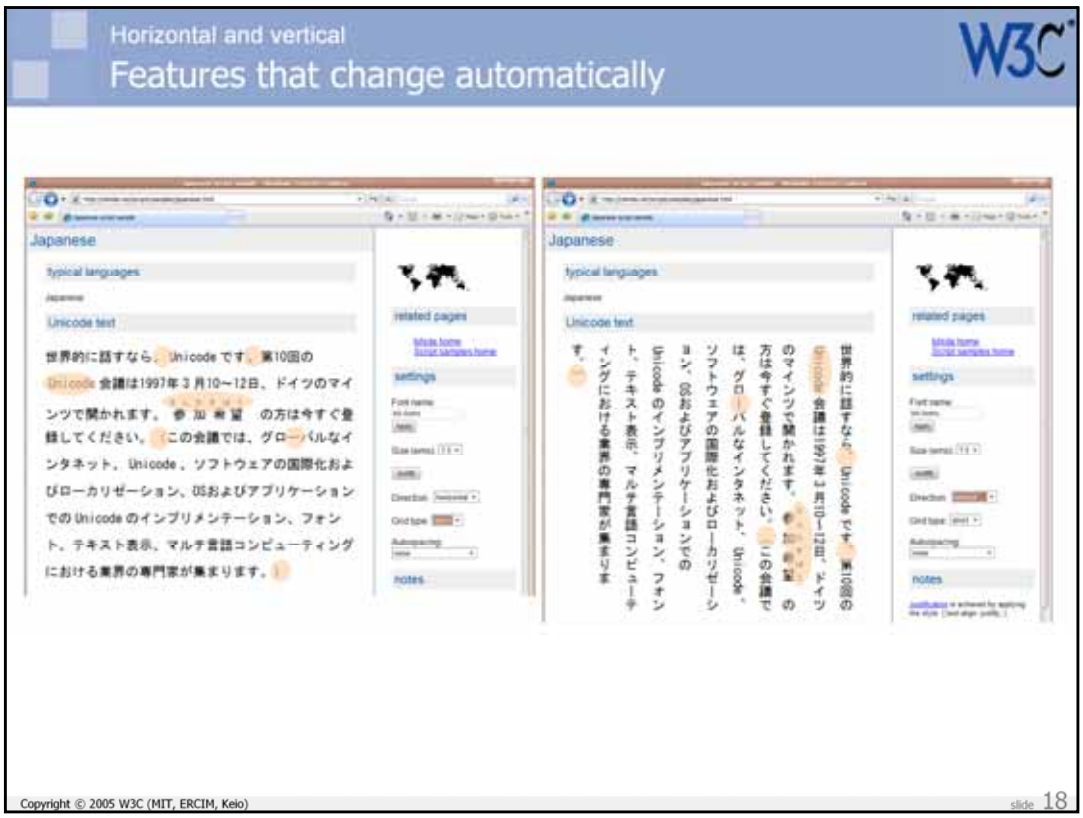
Figures and surrounding text tend to be positioned relative to the kihon hanmen grid.


Horizontal and vertical
W3C
Converting between horizontally and vertically set text
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 16

You may think that you would be able to switch between horizontal and vertical text at the flick of a CSS switch. I even have a page that allows you to do that, using some JavaScript that changes the writing-mode property on a block of Japanese text (see http://rishida.net/scripts/samples/japanese.html - requires Internet Explorer 5.5+).

There are some differences between text in vertical and horizontal layouts that can be dealt with automatically, but there are others that are not so easy to handle.

The change shown on the slide is brought about by simply changing the value of the writing-mode property on the enclosing block from lr-tb (left-to-right block flow, top-to-bottom line progression) to tb-rl (top-to-bottom block flow, right-to-left line progression).



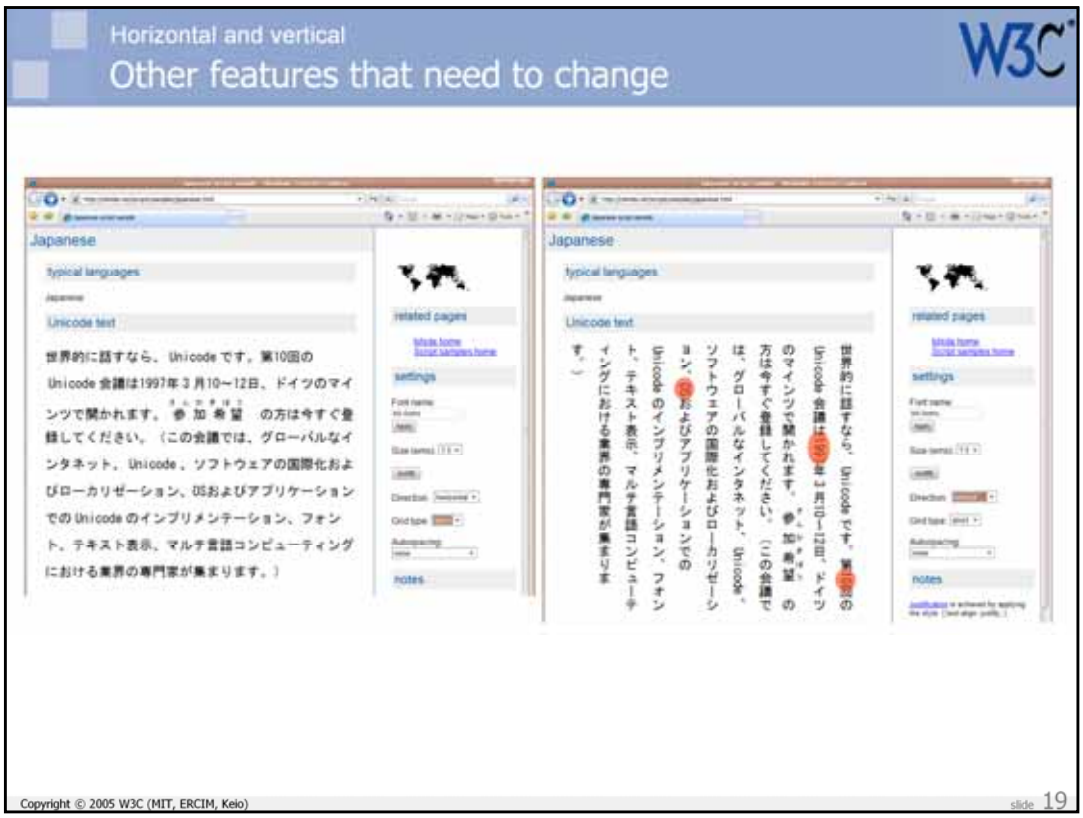
This slide shows some things that already work in Internet Explorer, using this simple switch.

The Japanese comma and full-stop move from the bottom left of the character square to the top right. This is achieved by choosing a vertical-text-specific glyph from the font. The character codes remain the same.

Sound elongation marks and punctuation such as parentheses are rotated.

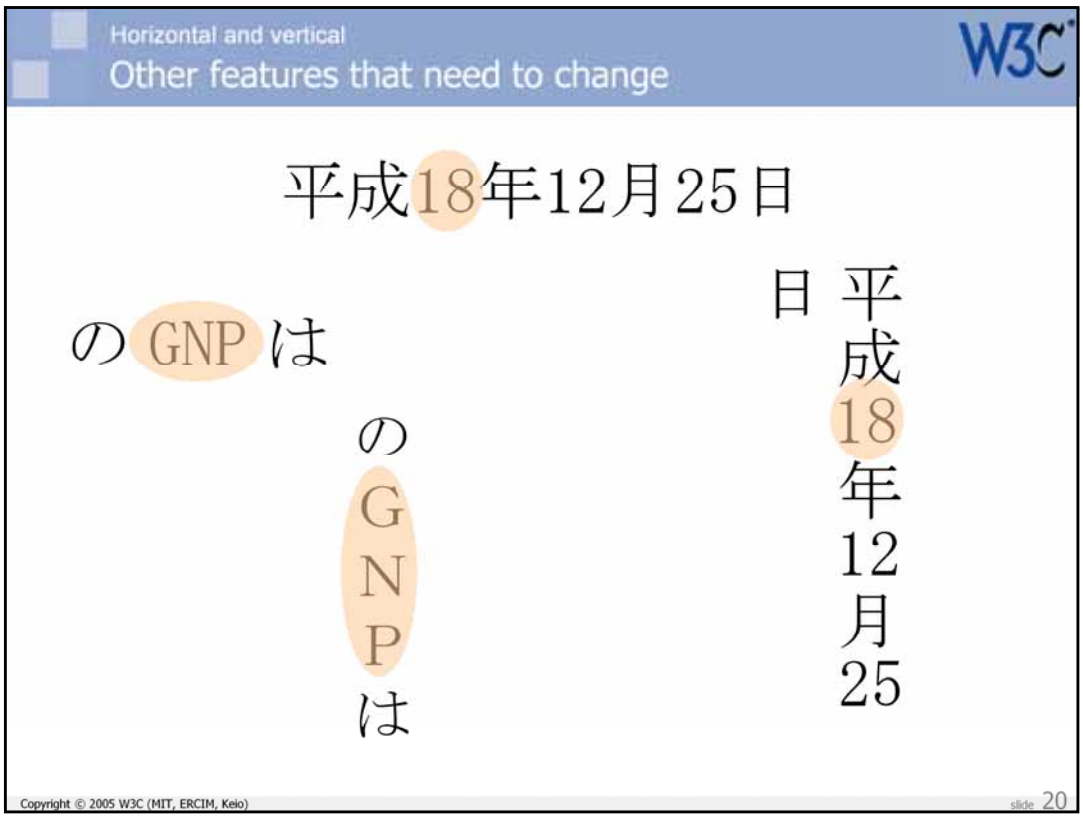
Runs of Latin text flow down the page in vertical lines.

Ruby text moves from above the horizontal line to the right of the vertical line.



Unfortunately, there are other aspects of the text that should have changed, but didn't. For example, some numbers would have looked better as kanji characters, rather than Western digits, in vertical text; other numbers should run horizontally within the vertical line progression; and some Latin text that is used as an acronym would look better if the characters were not rotated.

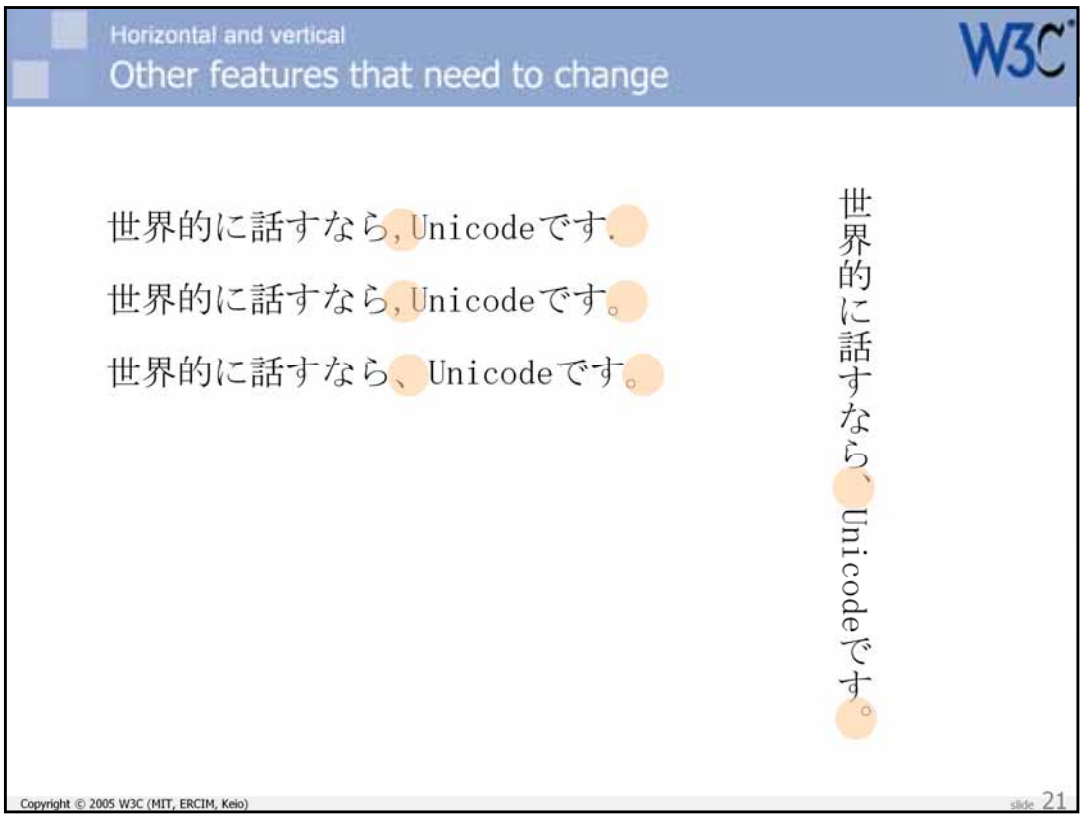
But there are also other differences. Some can be worked around, others are less tractable. Let's take a look at some of the common differences between horizontally and vertically set Japanese text.



A Latin-script acronym or single letter is typically rendered without rotation in vertical text. It also uses a full-width shape, whereas horizontal text will typically use proportionally-spaced characters.

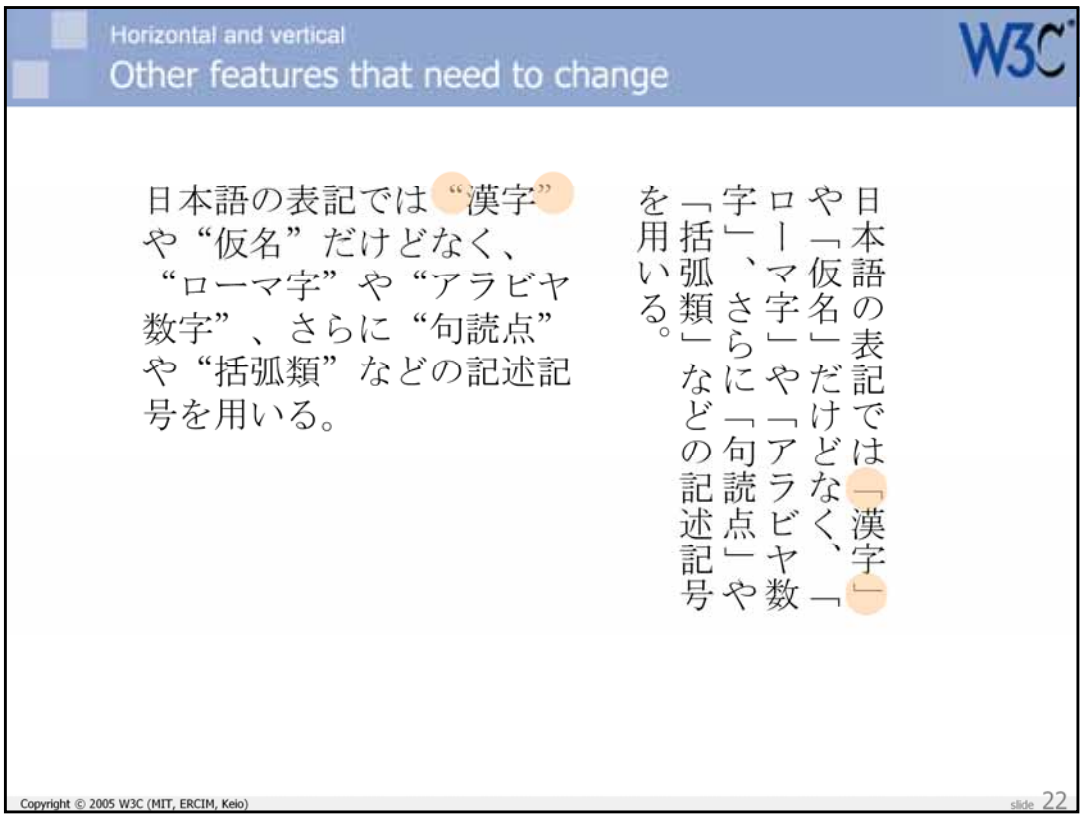
Short two-digit numbers are often horizontally set within the vertical line (this is called tate-chuu-yoko).

In both of these cases, you would need to have markup in the horizontal text to automatically achieve the desired result when the switch to vertical is flipped.



Punctuation characters tend to be used differently in horizontal and vertical text.

Due to Western influence, Japanese authors sometimes use Western commas and punctuation in text. Vertical text, on the other hand, should only use the ideographic comma and full stop.

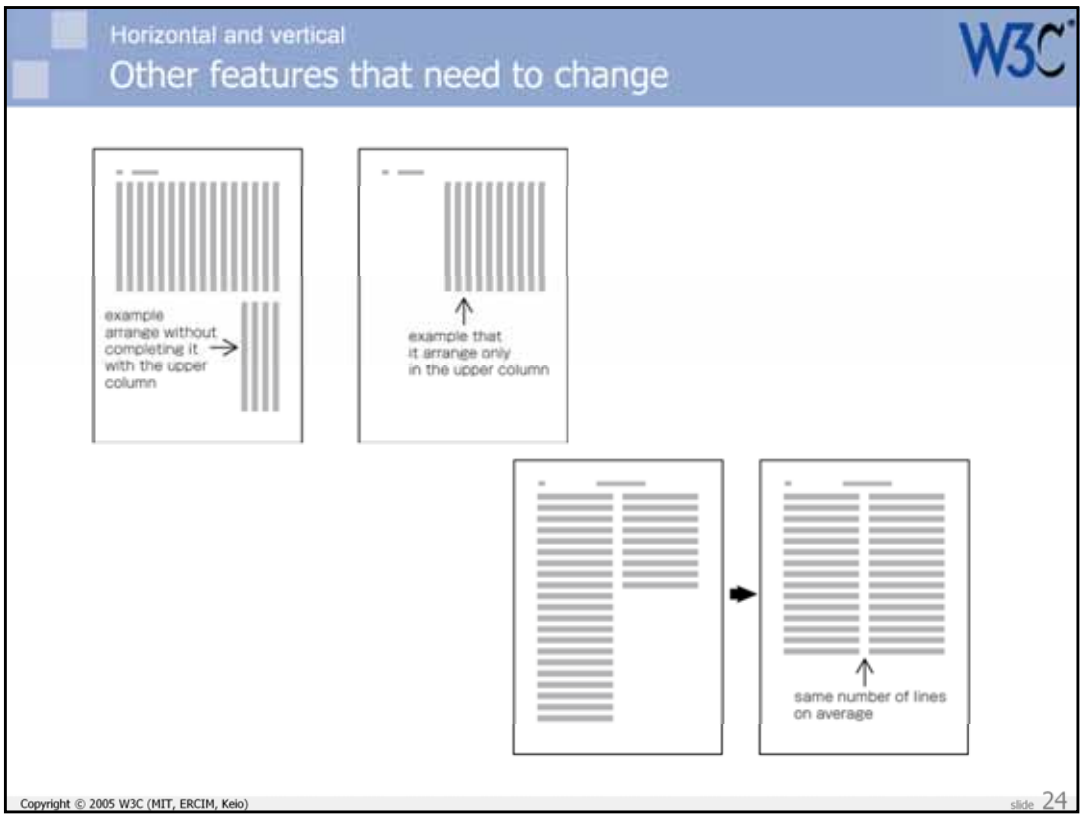


Similarly, there are some differences in the usage of quotation marks in horizontal vs. vertical text. For example, horizontal text often uses double or single quotation marks as an alternative to left and right corner brackets. Vertical text, on the other hand, doesn't use the quotation marks (except sometimes for embedded Latin runs); it uses corner brackets or double prime marks. Double prime marks are not used for horizontal text.

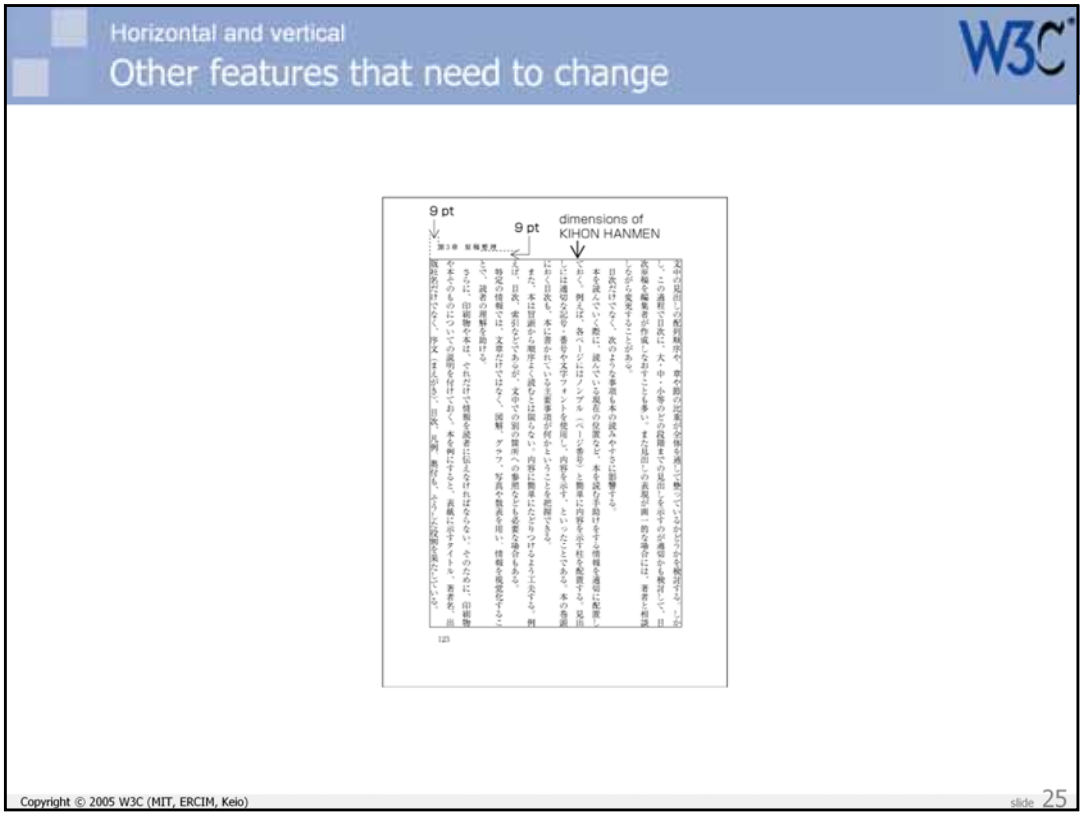
Similar rules apply to parentheses. Left and right tortoise shell brackets (﹝ ﹞) are a vertical equivalent of left and right square bracket ([ ]). Square brackets should be used in horizontal text except for special cases.

Ideographic numerals were traditionally used in vertically set text, instead of Western-Arabic numerals. (Road numbers and car numbers are examples of exceptions). However, newspapers and other publications have been adopting Western-Arabic numerals more in vertical writing mode.

In vertically set text, symbols for units are usually described with katakana characters, eg. センチメートル(centimeter) or センチ(abbreviated form for centimeter). In horizontally set text, the International System of Units(SI) is usually used, eg. "cm".

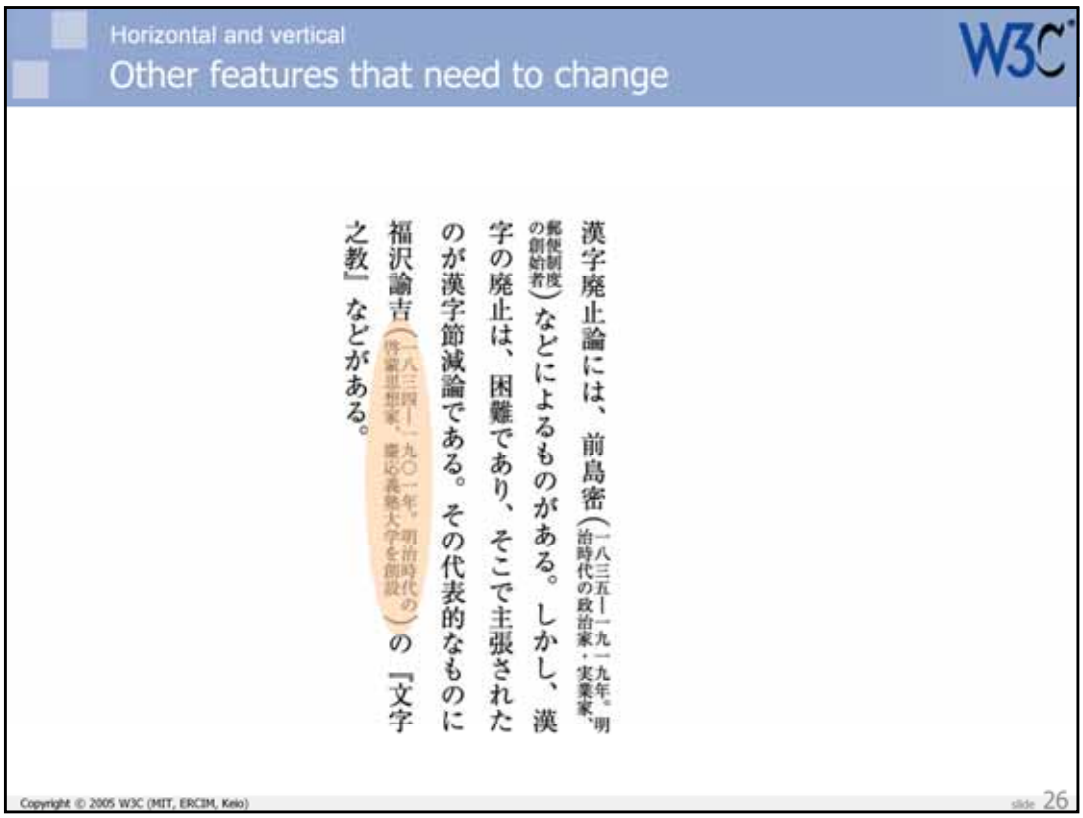


Where a new page is started and the current page isn't full of text, there may be different strategies for balancing columns on the current page. Typically, text columns in vertically set text end without any balancing. Columns in horizontally set text, however, are typically balanced so that the two columns are the same length.

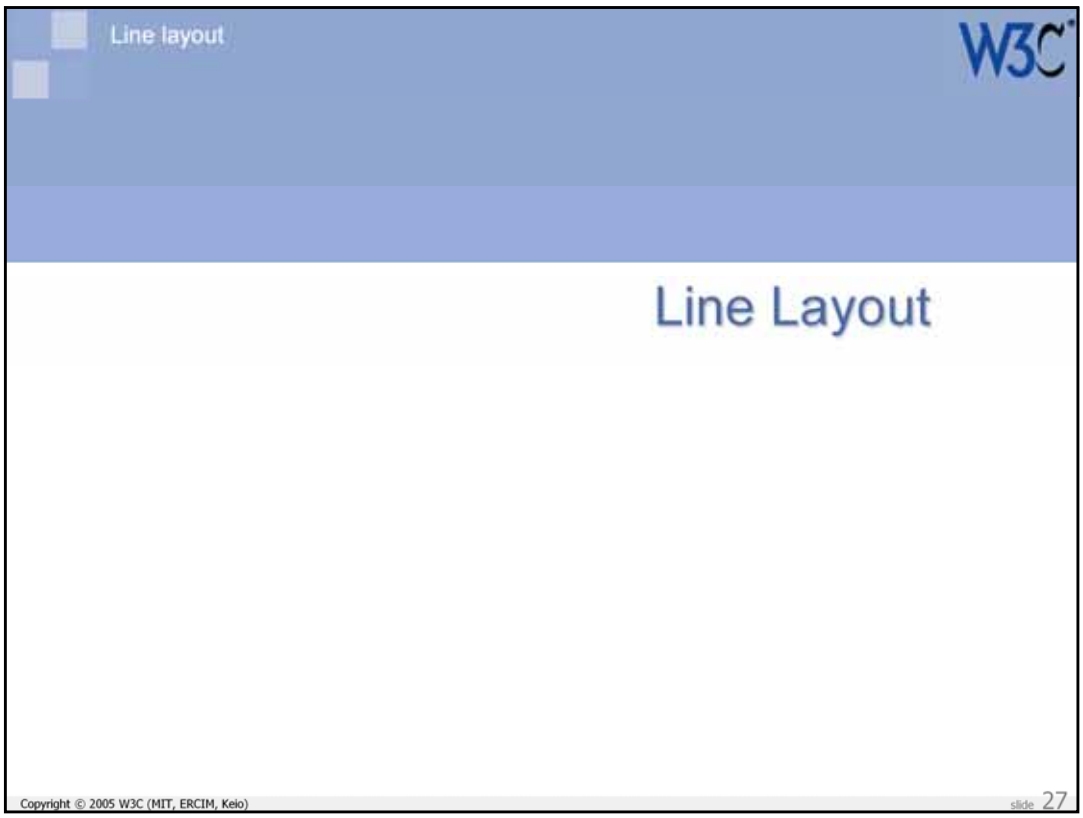


There are also differences in page layout details between horizontally and vertically set pages that are not automatically achieved.

For example, when positioning running heads and page numbers horizontally with reference to kihon hanmen in vertical writing mode, the amount of vertical space between the edge of kihon hanmen and the running head is one full width character size based on the kihon hanmen character sizes. If kihon hanmen is horizontally set, you need to take more vertical space than the character size in kihon hanmen.



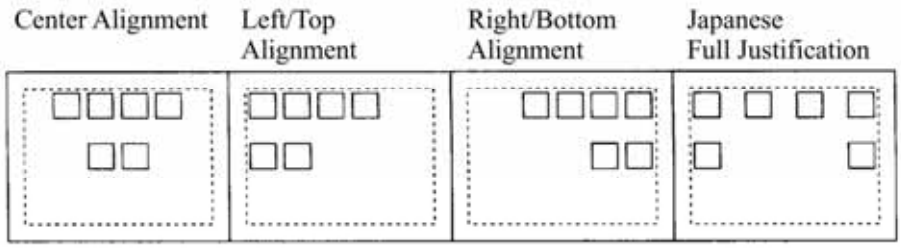
Typographic conventions such as warichu are much more common in vertically set text than horizontal.


Line layout
W3C
Line Layout
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 27

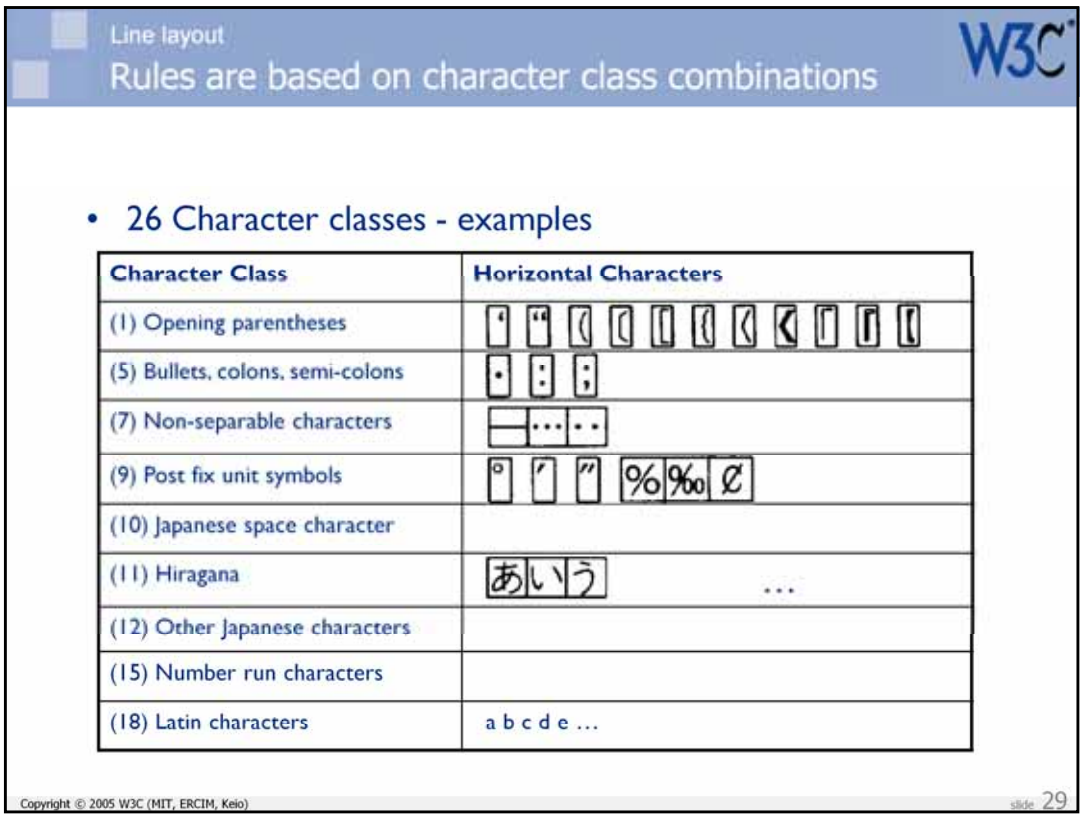
Line layout

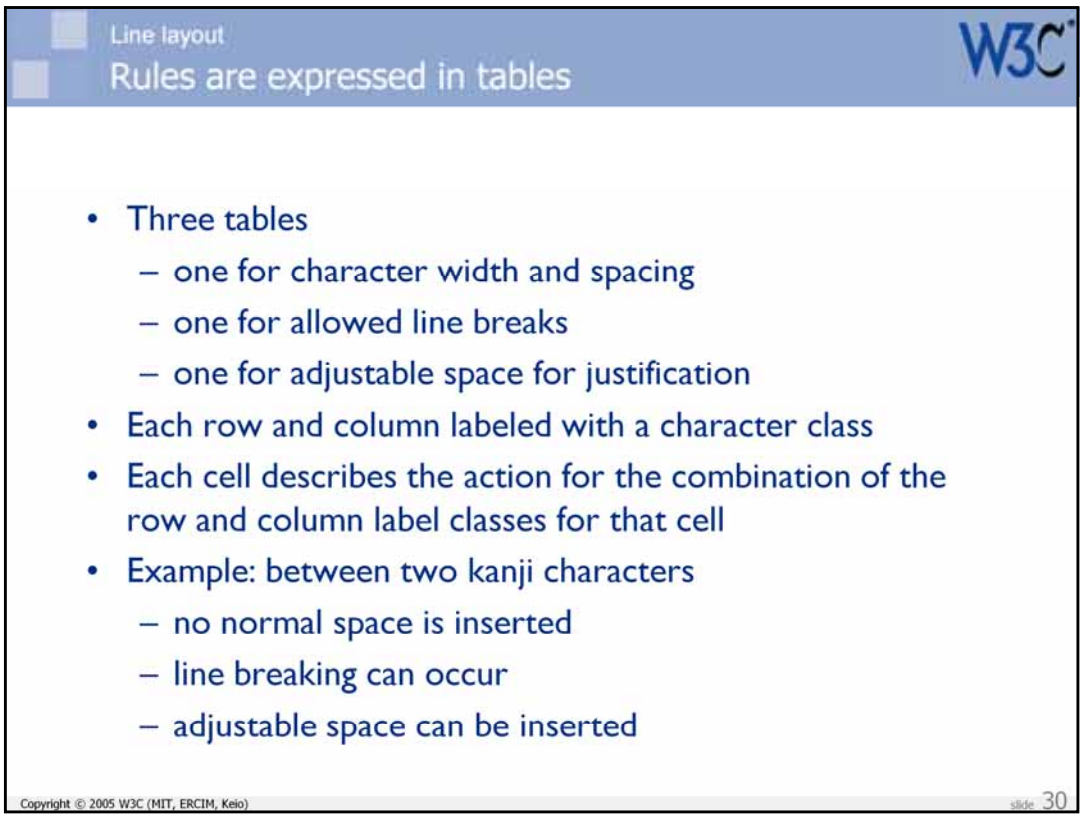
Rules for line layout in Japanese

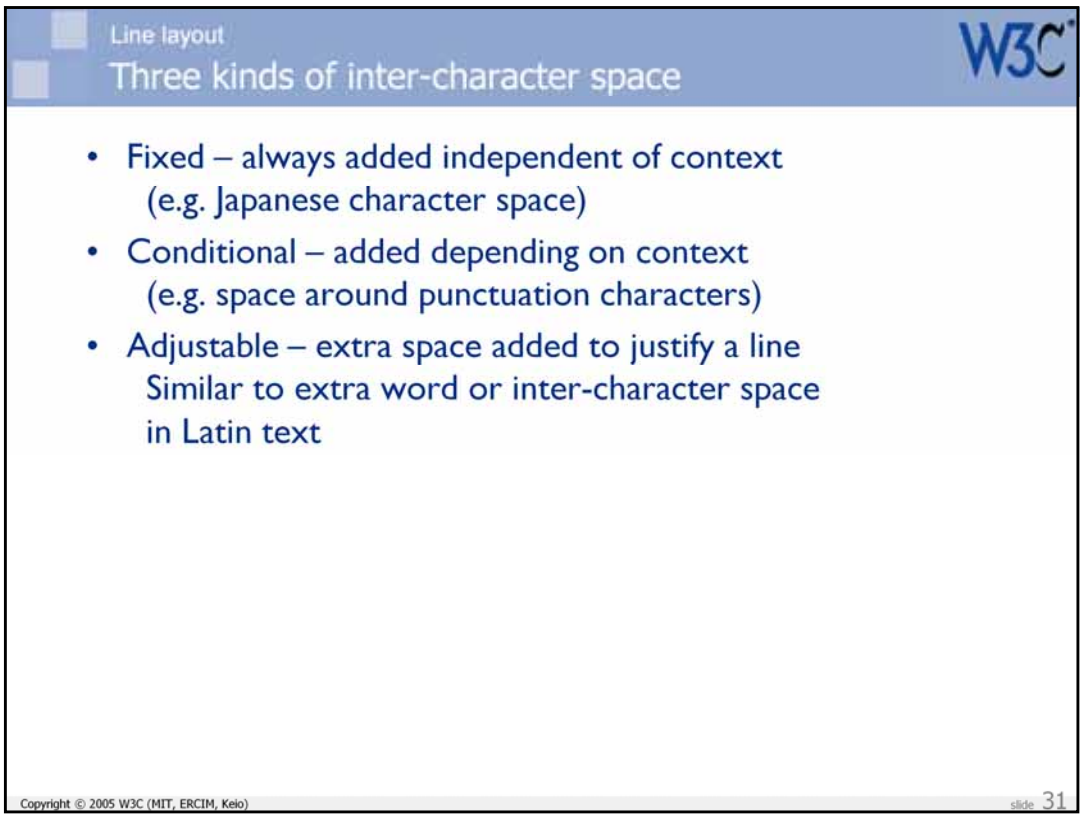
- Three basic sets of rules
  - for spaces between characters
  - for places at which line breaks are allowed
  - for justifying a line by adding adjustable space

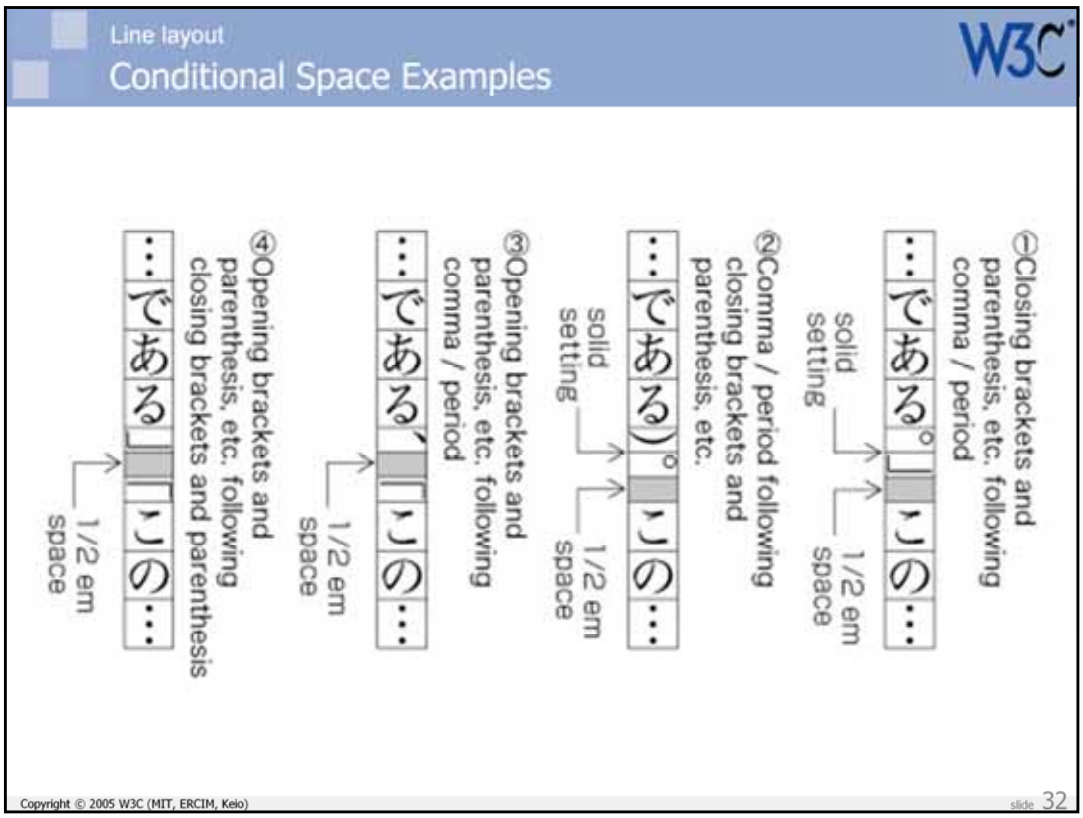


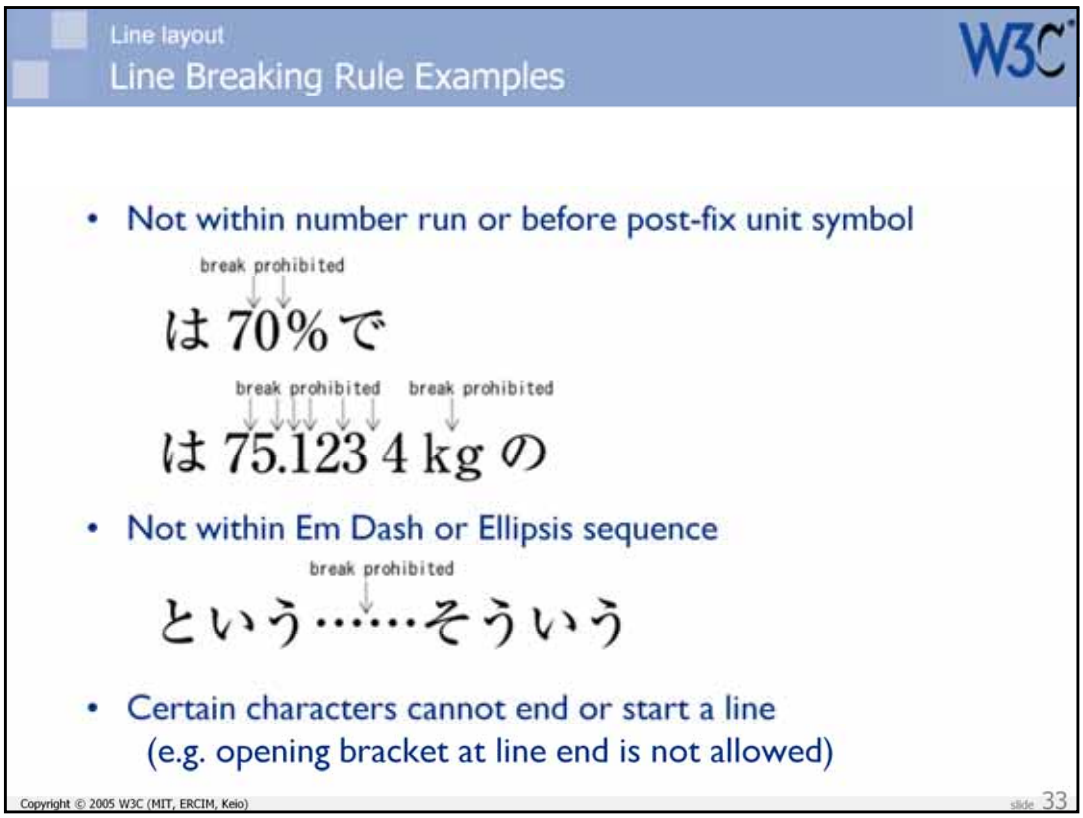
Copyright © 2005 W3C (MIT, ERCIM, Keio)

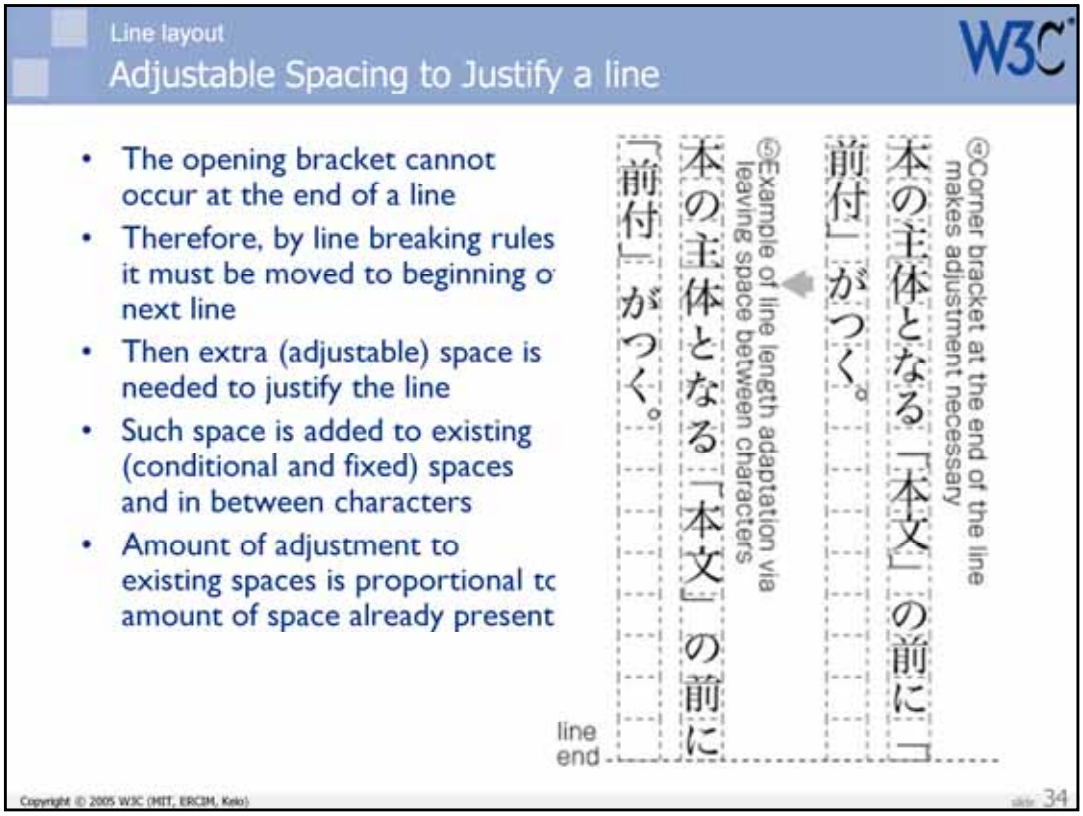

Line layout
Rules are based on character class combinations
W3C
• 26 Character classes - examples
Character Class
Horizontal Characters
(1) Opening parentheses
(5) Bullets, colons, semi-colons
(7) Non-separable characters
(9) Post fix unit symbols
(10) Japanese space character
(11) Hiragana
あいう
...
(12) Other Japanese characters
(15) Number run characters
(18) Latin characters
a b c d e ...
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 29

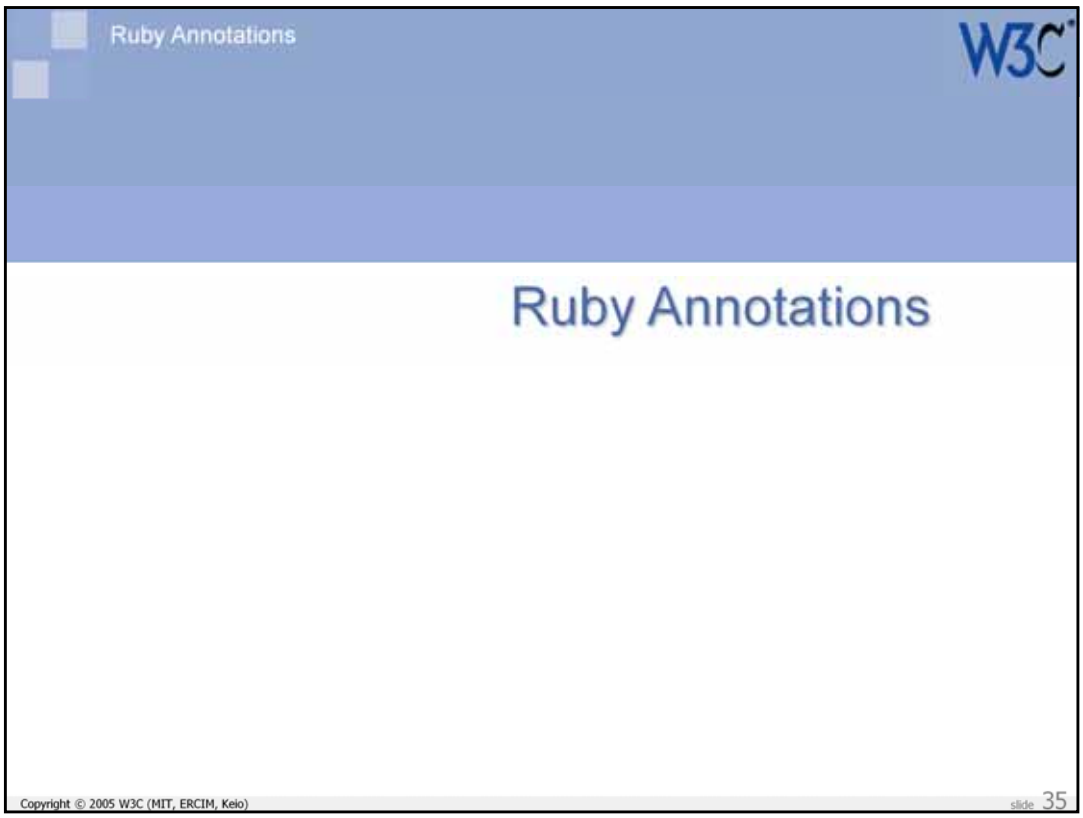

Line layout
Rules are expressed in tables
W3C
• Three tables
– one for character width and spacing
– one for allowed line breaks
– one for adjustable space for justification
• Each row and column labeled with a character class
• Each cell describes the action for the combination of the row and column label classes for that cell
• Example: between two kanji characters
– no normal space is inserted
– line breaking can occur
– adjustable space can be inserted
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 30


Line layout
Three kinds of inter-character space
W3C
• Fixed – always added independent of context
(e.g. Japanese character space)
• Conditional – added depending on context
(e.g. space around punctuation characters)
• Adjustable – extra space added to justify a line
Similar to extra word or inter-character space
in Latin text
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 31

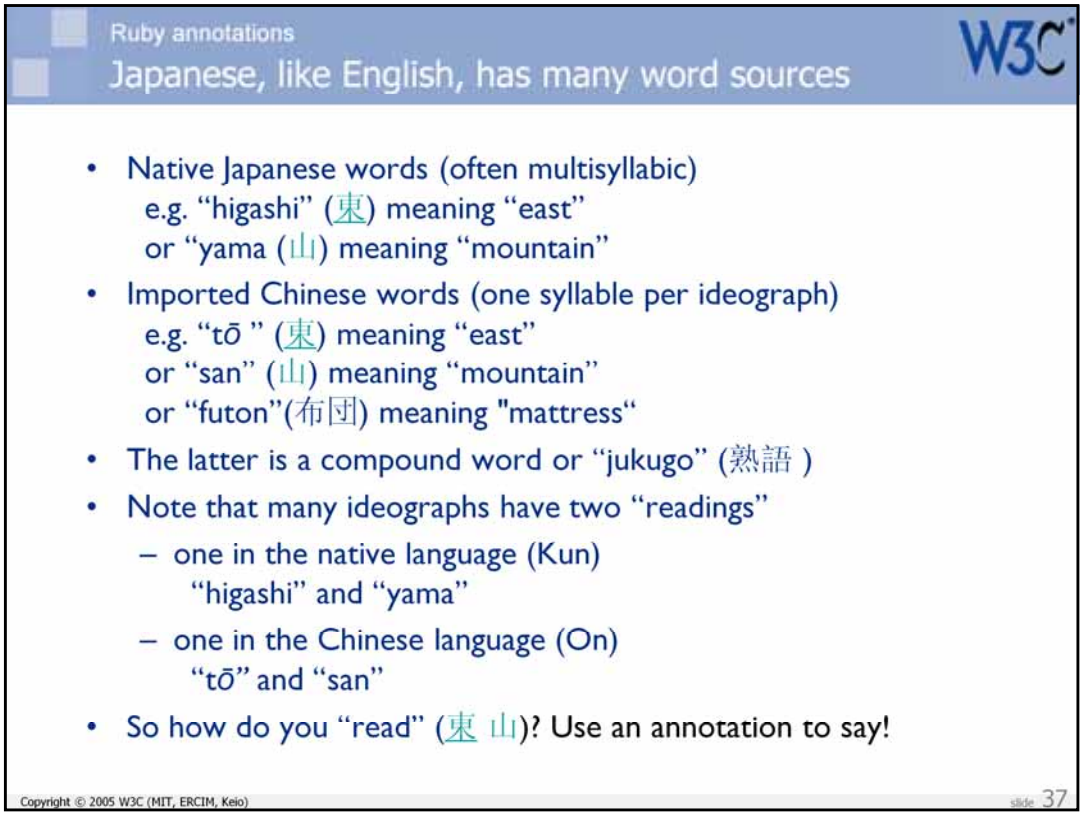

Line layout
Conditional Space Examples
①Closing brackets and parenthesis, etc. following comma / period
…である。」「トリの…
solid setting
1/2 em space
②Comma / period following closing brackets and parenthesis, etc.
…である)。「トリの…
solid setting
1/2 em space
③Opening brackets and parenthesis, etc. following comma / period
…である、「トリの…
1/2 em space
④Opening brackets and parenthesis, etc. following closing brackets and parenthesis
…である」「トリの…
1/2 em space
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 32


Line layout
Line Breaking Rule Examples
W3C
• Not within number run or before post-fix unit symbol
break prohibited
は 70%で
break prohibited
break prohibited
は 75.123 4 kg の
• Not within Em Dash or Ellipsis sequence
break prohibited
という……そういう
• Certain characters cannot end or start a line
(e.g. opening bracket at line end is not allowed)
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 33


Line layout
Adjustable Spacing to Justify a line
W3C
The opening bracket cannot occur at the end of a line
Therefore, by line breaking rules it must be moved to beginning o next line
Then extra (adjustable) space is needed to justify the line
Such space is added to existing (conditional and fixed) spaces and in between characters
Amount of adjustment to existing spaces is proportional tc amount of space already present
ⓒCorner bracket at the end of the line makes adjustment necessary
本の主体となる「本文」の前に「
前付」がつく。
ⓔExample of line length adaptation via leaving space between characters
本の主体となる「本文」の前に
「前付」がつく。
line end
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 34


Ruby Annotations
W3C
Ruby Annotations
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 35

Japanese is a multisyllabic language that is very different that mono syllabic, tonal Chinese, but lacking a script of their own, the Japanese adopted the Chinese ideographic script to write Japanese. When this proved not adequate to the task they added two phonetic scripts (first hiragana and then katakana). Today, hiragana is used to express things like verb and adjective inflections (endings) (lacking in chinese) and to write words that lacked Chinese character(s). And, katakana is used to express words imported in the 19th century or later and for proper names and/or emphasis. Finally, the latin alphabet is used to write Japanese in a form that foreigners might understand.



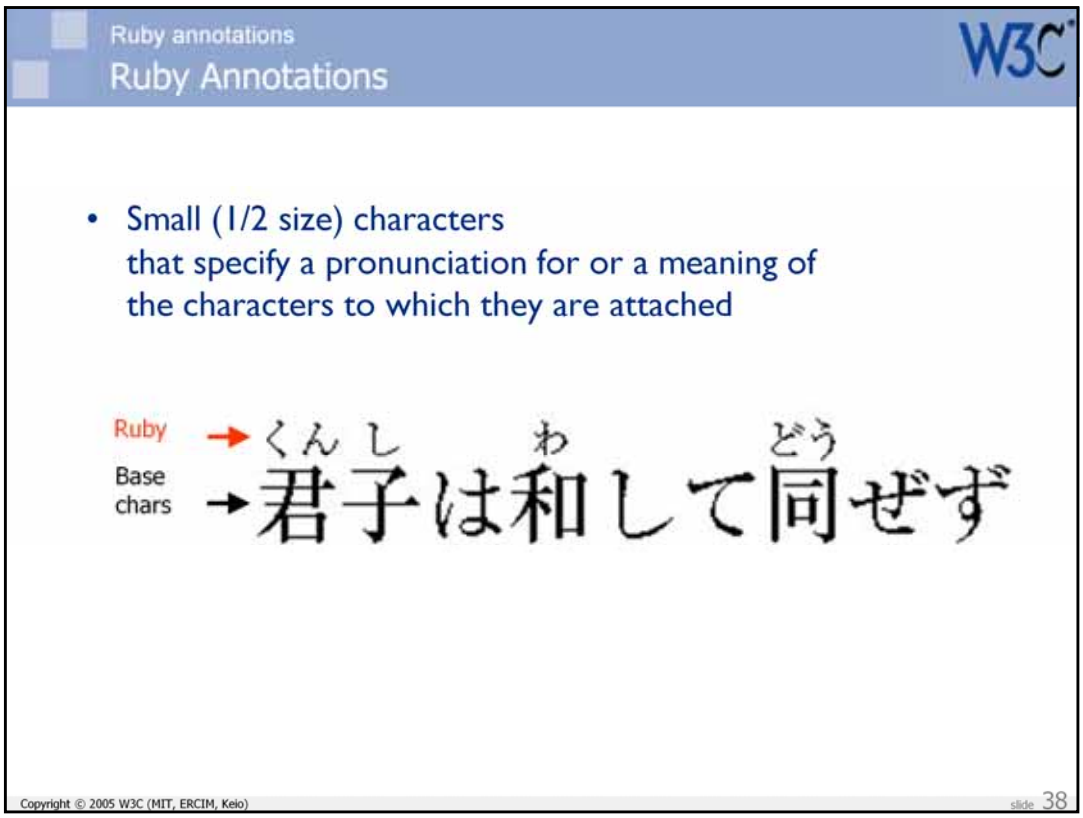
English is composed of words derived from Anglo-Saxon (pig, sheep) and Norman French (pork, mutton). Similarly, Japanese has words from the native spoken langauge, words imported from China, and words imported from Western (or other non-chinese) countries.

The native words are typically multi-syllabic and without tones. Typically a single kanji is assigned to each Japanese word, independently of how many syllables is has. So we have "higashi" (3 syllables) for "east" and "yama" (2 syllables) for "mountain". Since at the time the Chinese script (Kanji) was adopted for writing Japanese, it typically had words, in Chinese, for all the things Japanese had words for, there are also Sino-Japanese pronunciations for the same kanji that represent Japanese words. Here there is "T*ō*" for "east" and "san" for "mountain. The Chinese "readings" of the kanji characters are normally mono-syllabic (but they have lost the associated Chinese tone).

Since Chinese was more sophisticated that Japanese at the time of borrowing of Kanji, there are words that have multiple kanji in their "spelling". There compound words are called "jukugo" in Japanese. The individuals characters in "futon": fu (character 737) means "spread" or "cloth" and "ton" (character 243) means "group or troupe". So the literal combination would be "group of cloth" which is an OK description of a "mattress" which is what "futon" means as a compound word.

Another example in Japanese is the pair **Shinto** (神道 *Shintō*) **meaning the "path/way of the gods" and**

The word *Shinto*, from the original Chinese *Shendao* (神道),[1] combines two *kanji*: "*shen*" (神 *"shen"*) **(Chinese loanwords usually retain their Chinese pronunciation, hence "shin" not "kami" the Japanese for "gods"), meaning gods or spirits; and "*dao*"** (道 *"dao"*)**, meaning a philosophical way or path.**



The ruby says, "ku" "n" "shi" (ha) "wa" (shi) (te) "do" "u" (ze) (zu)
where the ruby is quoted and the base kana is in parentheses

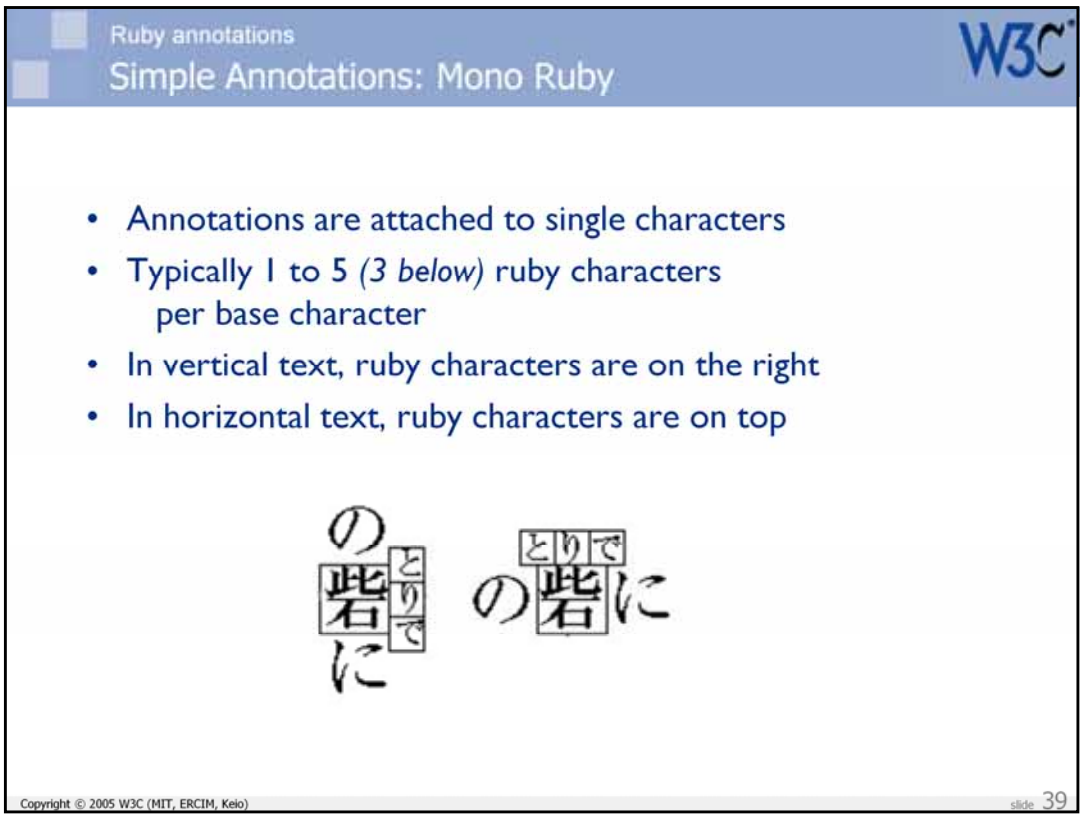
'kun' (character 479) is an honorific

'shi' (character 61) means child or person

so together they mean "wise man", "man of virtue or high rank"

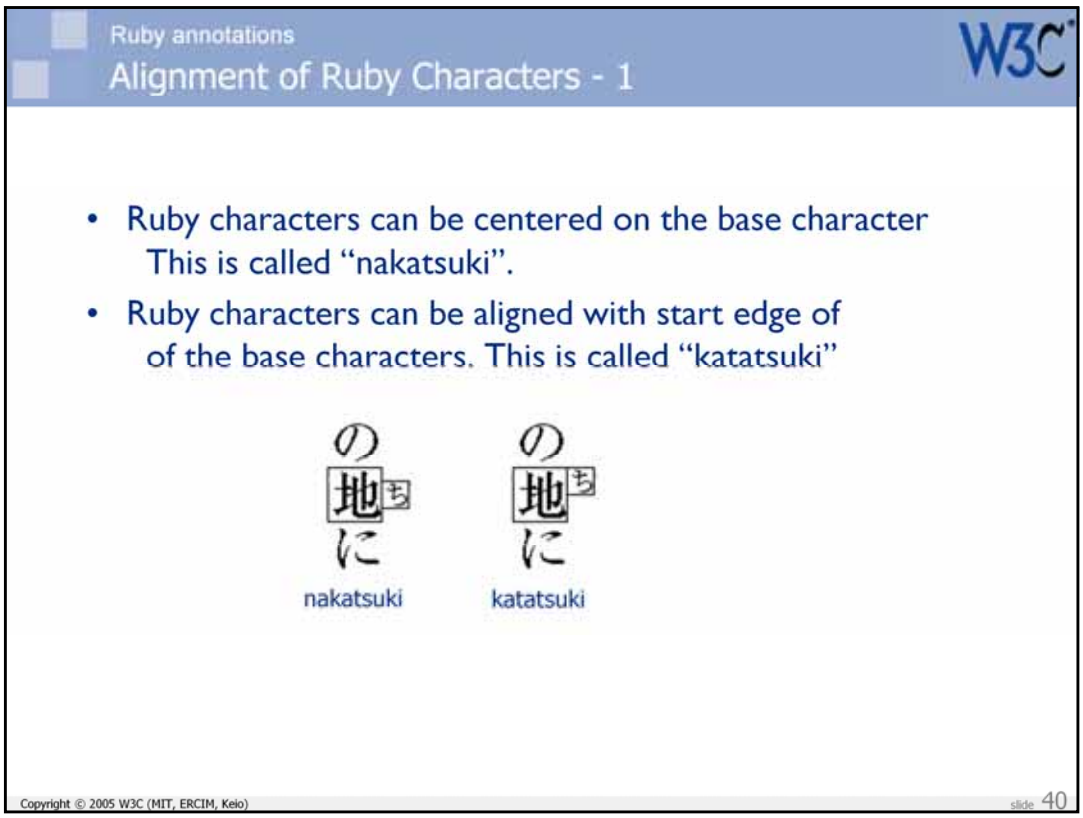
"wa" (character 277) means 'peace', "harmony", or 'Japan"

"do u" (character 185) seems to mean "same" (but normally without the "u" part)



The imaginary box (in which the character is centered) is drawn to more clearly show the positioning of the ruby characters relative to the base character.

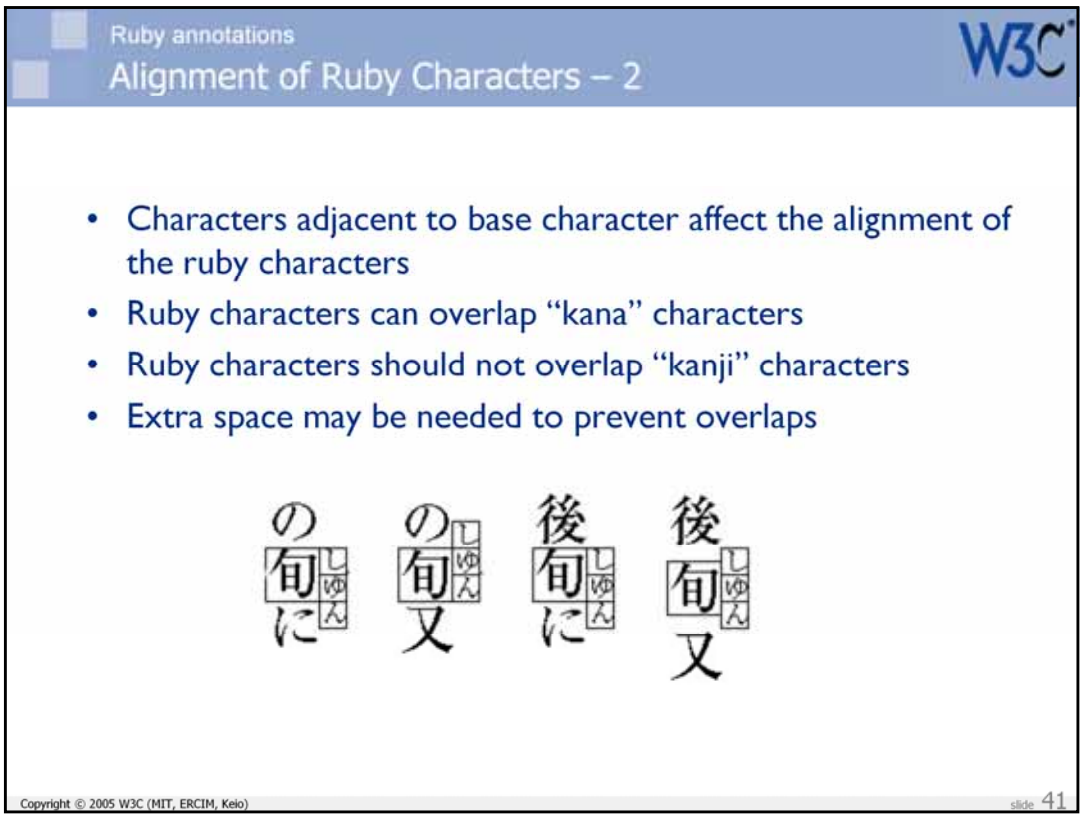
The hiragana seems to say, "to n de"



There are two approaches to positioning the ruby characters with respect to the base character. If there is an odd number of characters, as is the case here, then the center character can be aligned with the center of the base characters. This is called, "nakatsuki" alignment.

Alternatively, the initial character of the sequence of ruby characters can be aligned with the start edge of the base characters. This is called, "katatsuki' alignment. This approach was more often used with (hot) metal type setting and with digital text layout, the nakatsuki approach is more popular.

Note that if there are two ruby characters attached to the base character, then both approaches give the same results as long as the size of the ruby characters are ½ the size of the base characters.



The ruby annotation is: "shi' "yu" "n" (?)

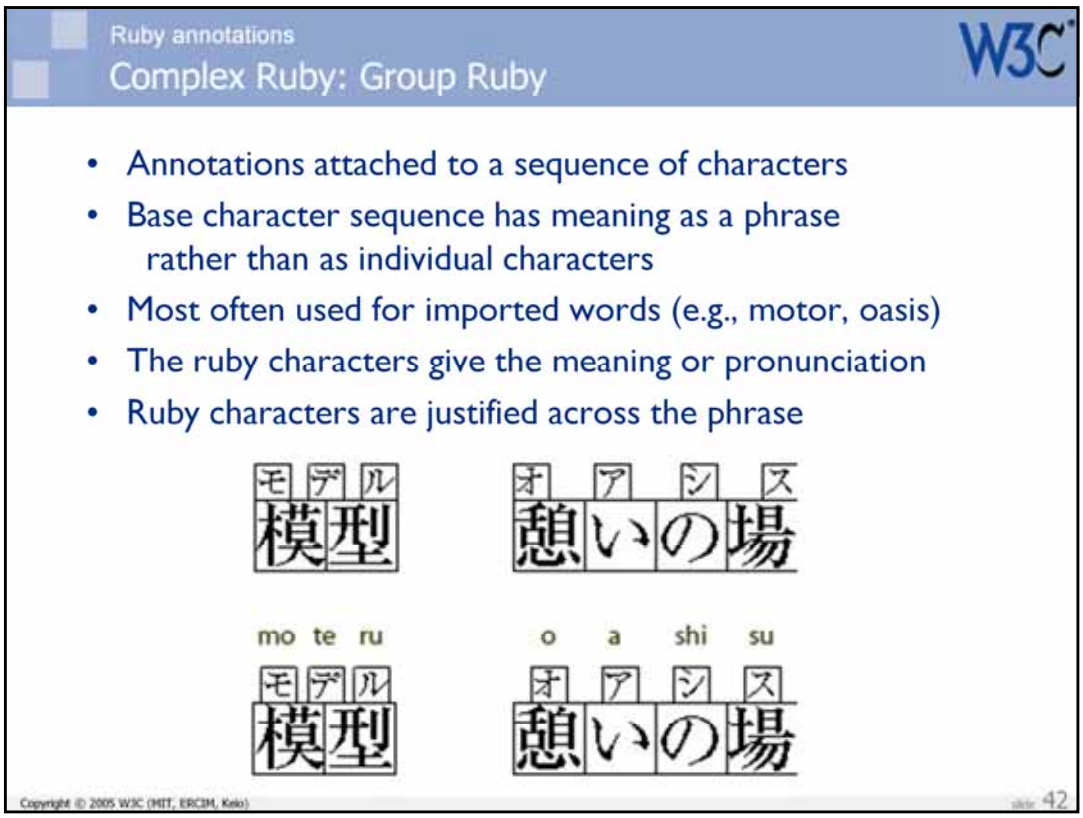
Going from left to right:

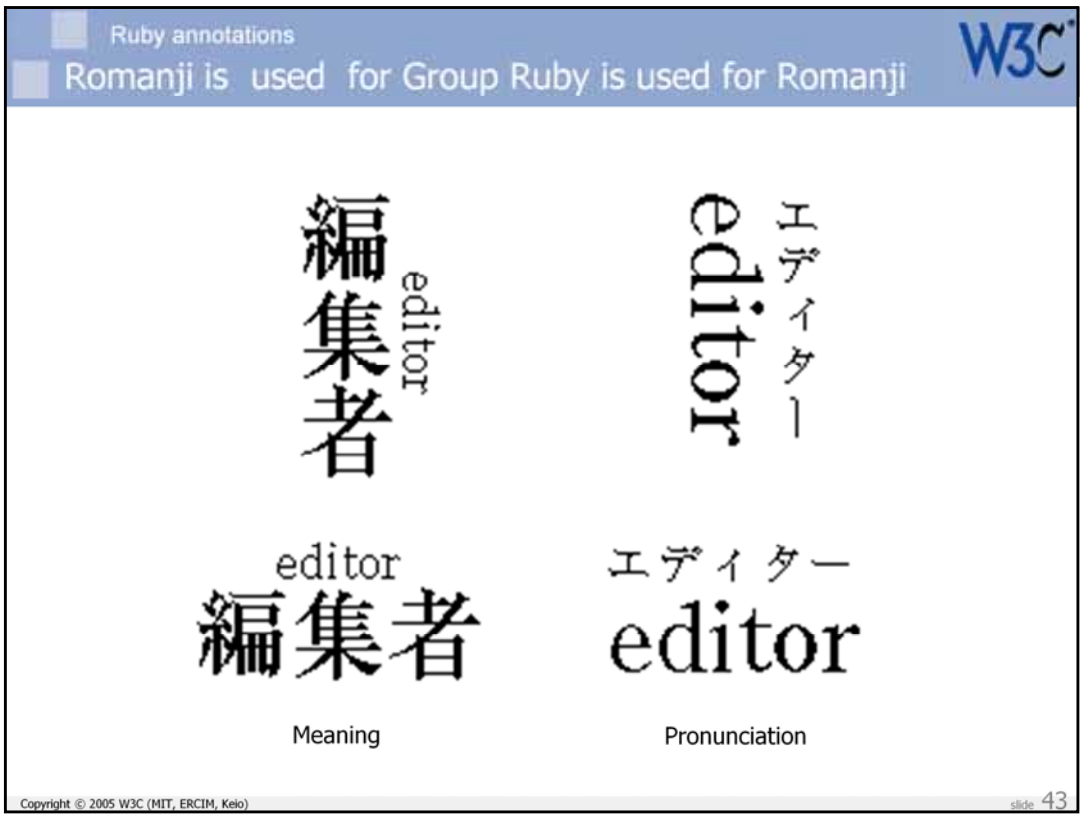
The first sequence has kana before and after the annotated kanji so overlap is allowed on either side, but the in katatsuki the first ruby character is normally aligned with the start edge of the base characters

The second sequence as a kanji character (hand) following the annotated kanji so the only allowable overlap is with the hiragana character that is before the annotated base character

The third sequence has a kanji before the annotated base character and kana afterwords, so the normal katatsuki rule works (actually is forced) in this case.

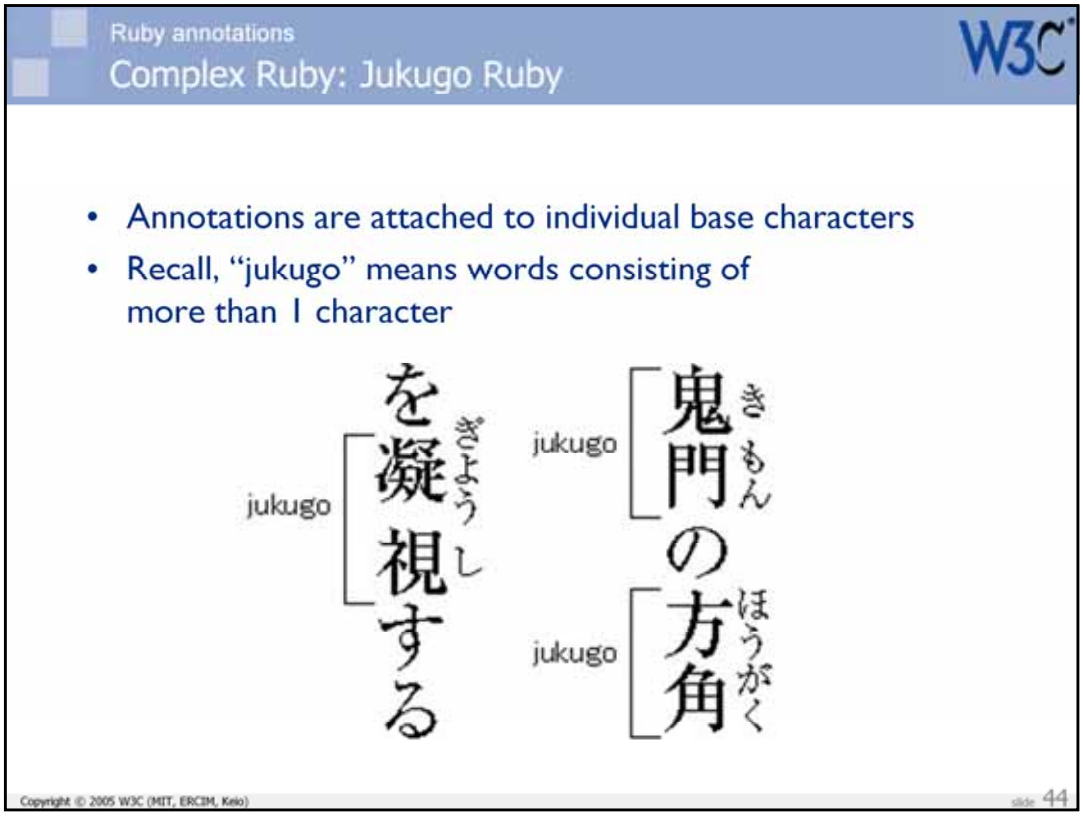
In the fourth and final sequence, there are kanji characters before and after the annotated base character and, therefore, it is necessary to (a) use nakatsuki alignment and (b) introduce space between the annotated base character and the kanji charaacters on either side. Using nakatsuki alignment means that the extra space is distributed equally on both sides of the annotated base character.


Ruby annotations
Complex Ruby: Group Ruby
W3C
• Annotations attached to a sequence of characters
• Base character sequence has meaning as a phrase rather than as individual characters
• Most often used for imported words (e.g., motor, oasis)
• The ruby characters give the meaning or pronunciation
• Ruby characters are justified across the phrase
モデル
模型
オアシス
憩いの場
mo te ru
o a shi su
モデル
模型
オアシス
憩いの場
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 42



In the lefthand set, the kanji is given an English meaning, "editor". Note that the latin characters are formatted/justified as the normally would be rather than spreading them across the whole "width" of the kanji base character sequence.

In the righthand set, the pronunciation of the English word, "editor" is given in katakana, which is evenly distributed across the "width" of the base (romanji) characters.

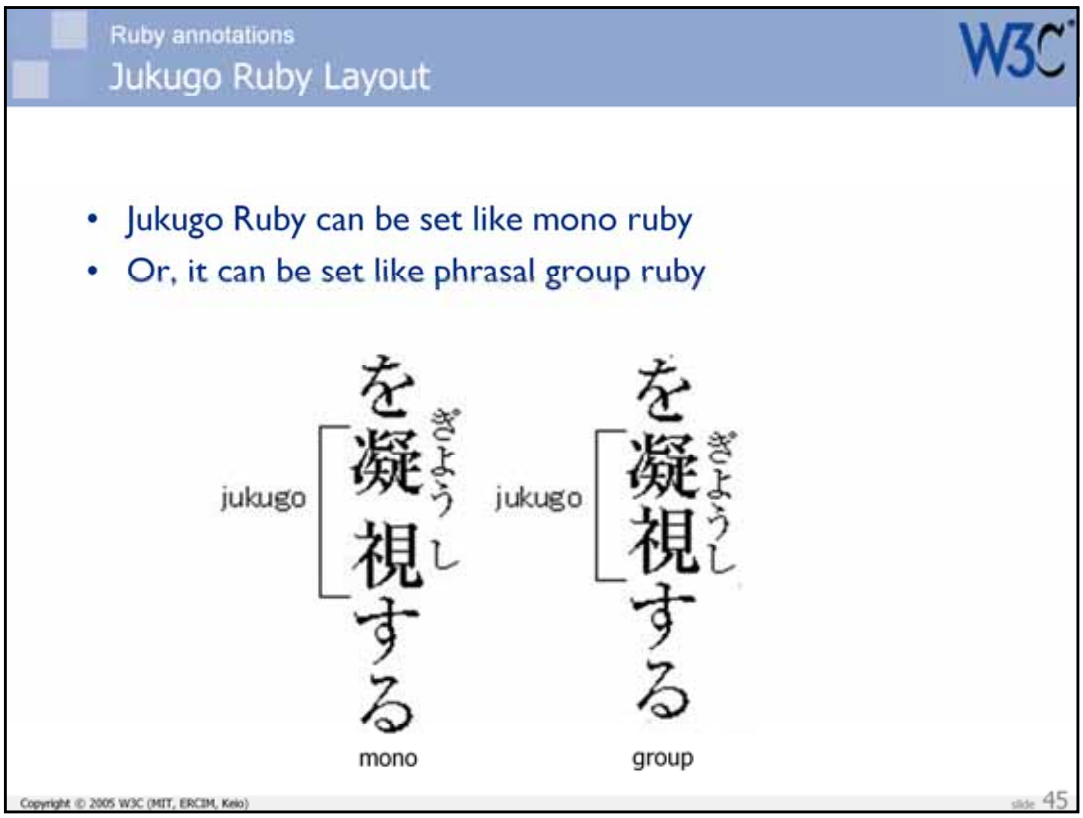


The example here is in two lines of Japanese and is read from right to left vertically

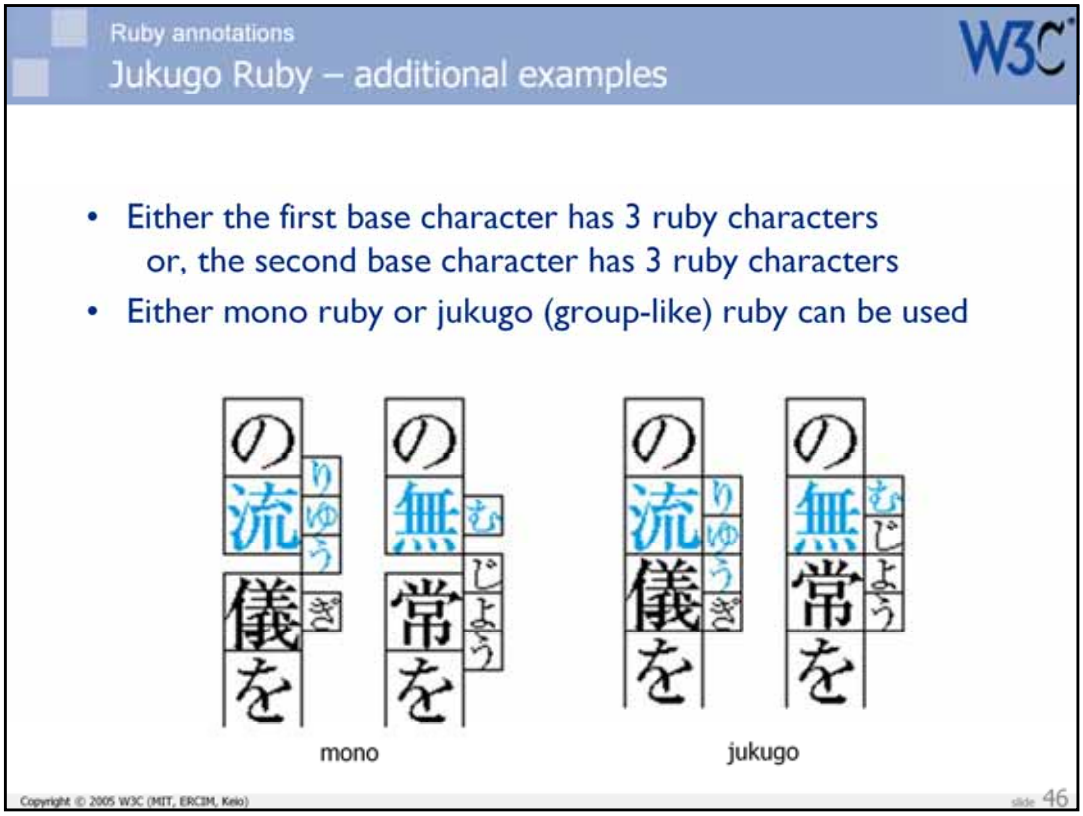
The first jukugo, the one on the top right, has only three ruby characters in its annotation: one on the first base charaacter and two on the second base character

The second jukugo, the one on the bottom right, has two ruby characaters on each of its two base characters

The third jukugo, the one on the left, has three ruby characters on its first base character,but only one ruby character on its second base character. Because the ruby cannot overlap a kanji character, there must be a space between the first and second kanji characters. But, this is somewhat ugly. so, next slide



In this slide, we have repeated the mono ruby setting of jukugo ruby, but show that a more beautiful setting results from treating the jukugo ruby as if it were group ruby.



Here are two examples of jukugo text with attached ruby annotations.

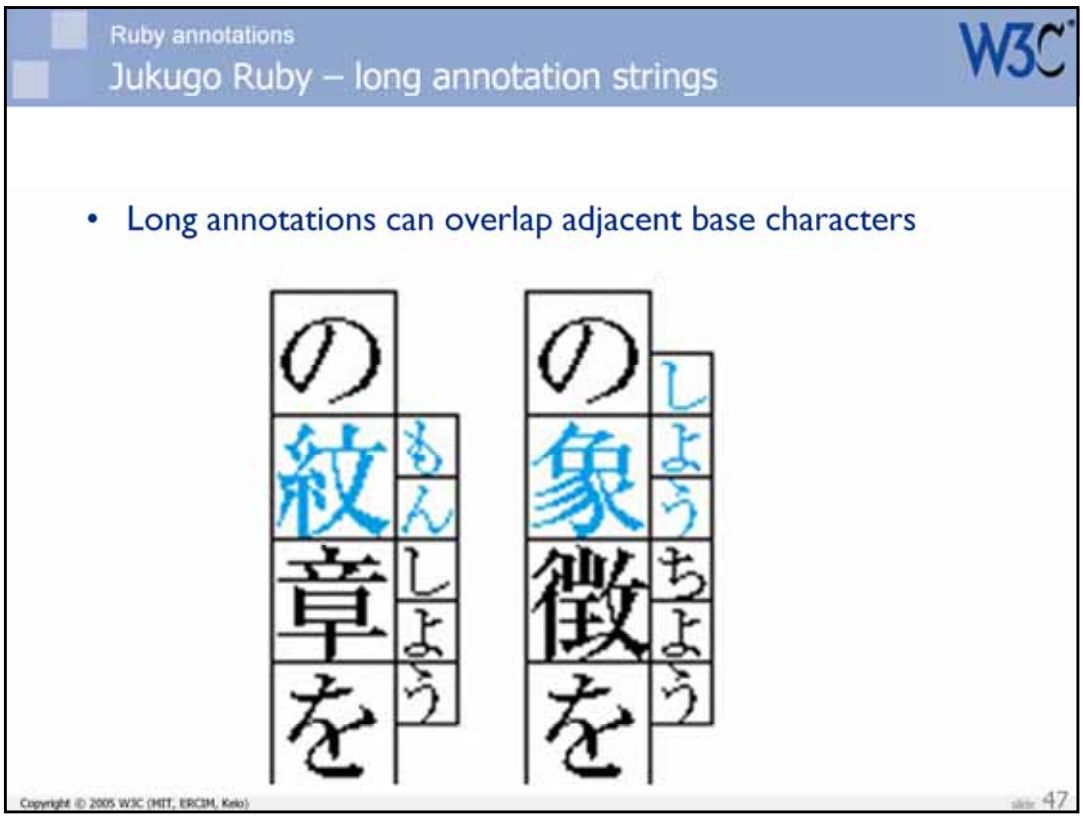
In the first example, the left one, the first base character has three attached ruby characters and the second base character only has one attached ruby character.

In the second example, the right one, the first base character has only one attached ruby character and the second base character only has three attached ruby character

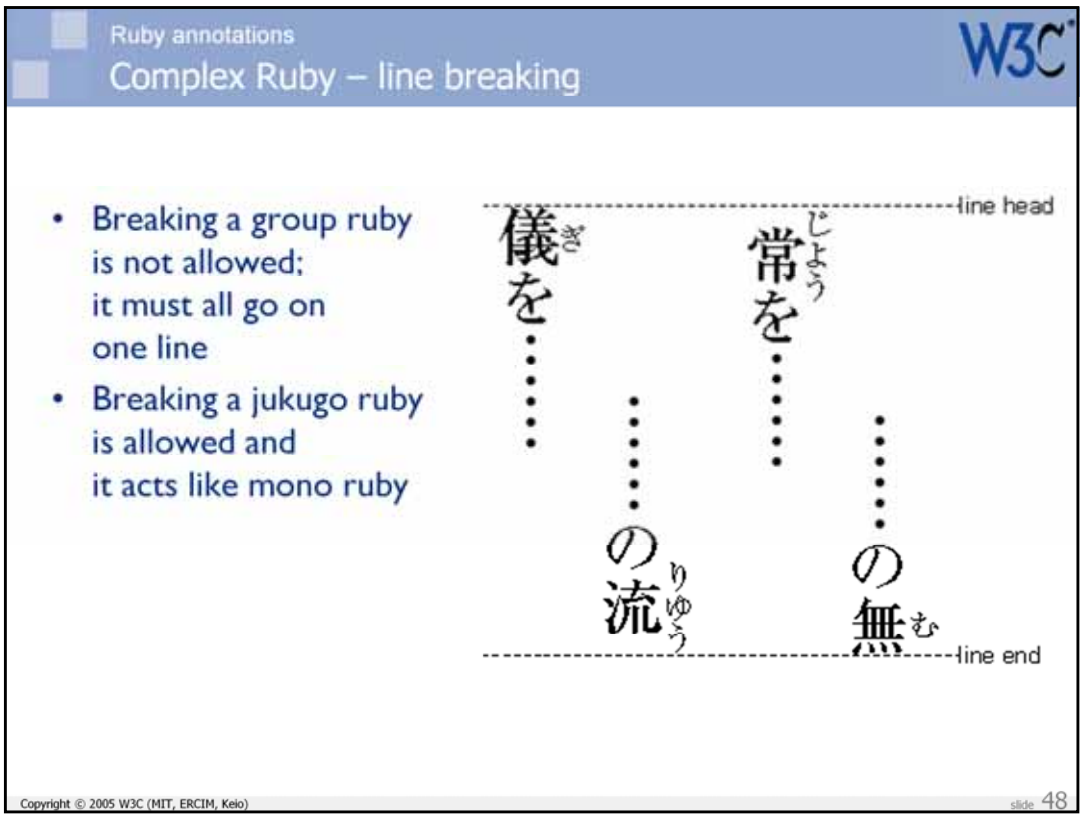
In both cases, the first base character and its attached ruby charaacters are shown in blue and the second base characters of the jukugo and its attached ruby characters are shown in black.

The pair of examples on the left show a setting of both examles using the rules for mono ruby. The second pair of examples show that the rules for jukugo ruby allow a much more pleasing setting of the text. The jukugo setting avoids breaking the Kihonhanmen grid. Note that it does not matter which character has 3 ruby characters attached as long as the other character has only one ruby attached.

So, what happens if the jukugo base characters have more that a total of 4 ruby characters attached?

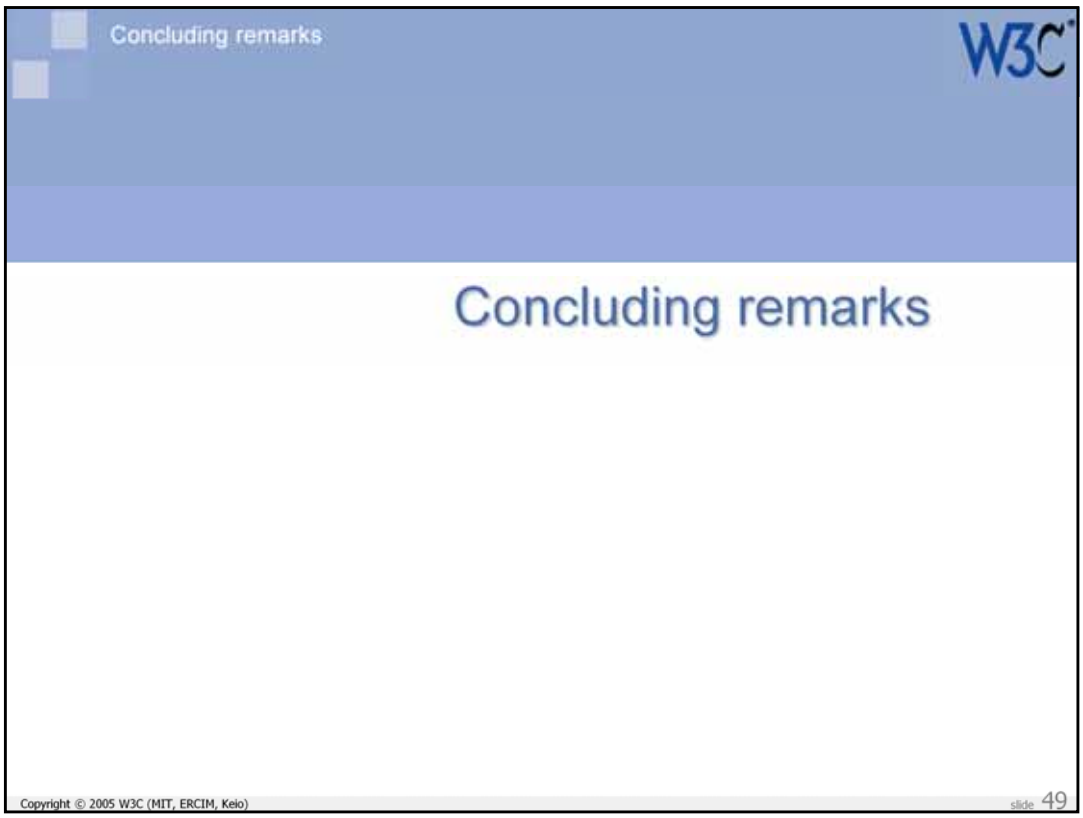


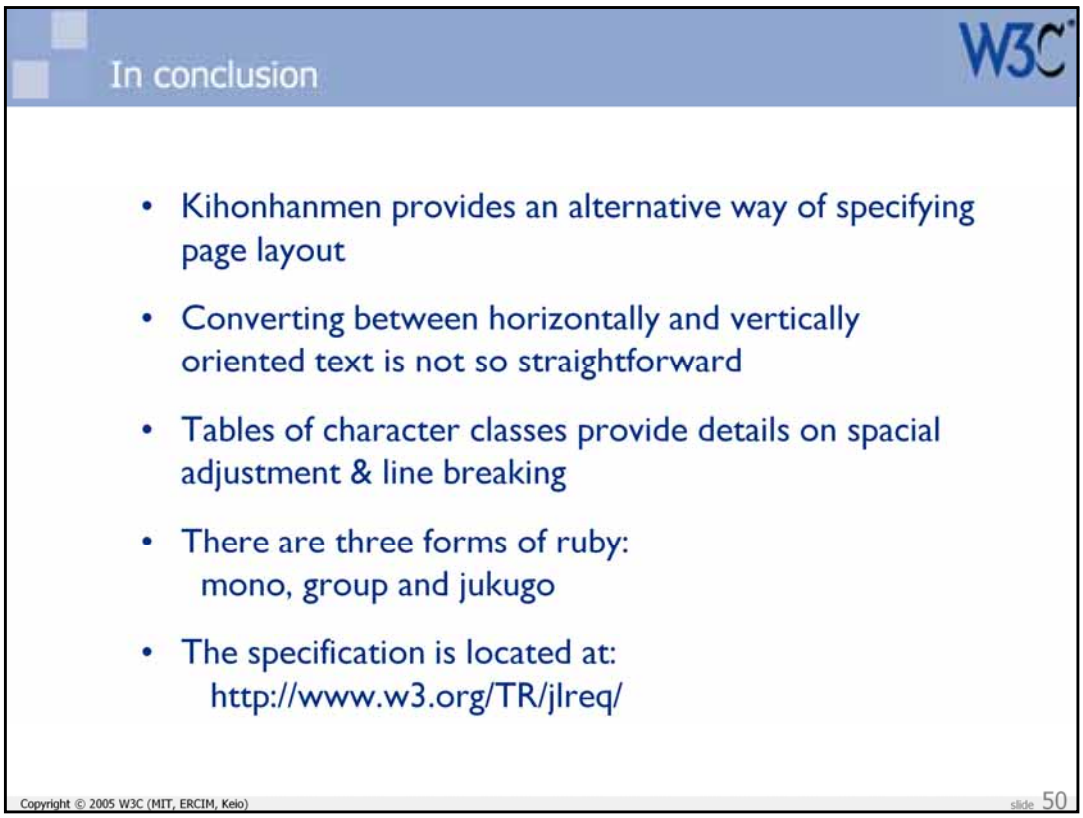
As for group or mono ruby, the overlaps are only allowed if the adjacent characters are not kanji; that is the case in these examples.

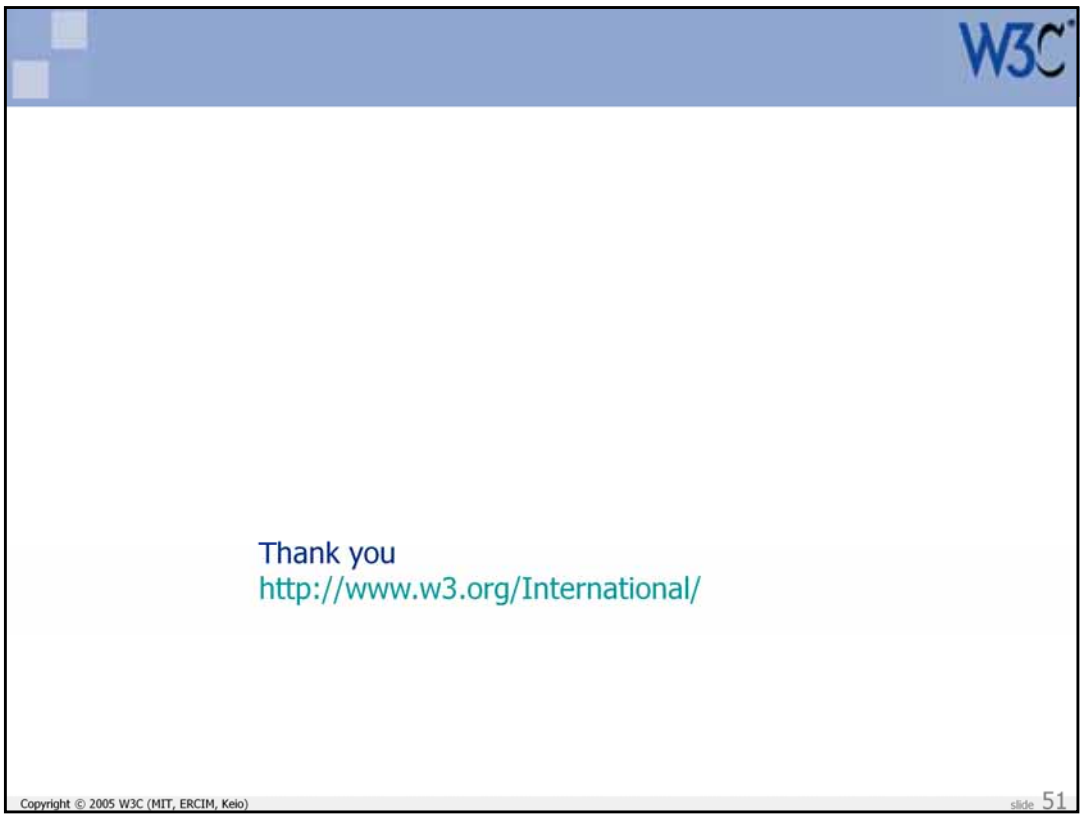


One major difference between jukugo and group ruby is that a line break may not occur within a group ruby sequence. For jukugo ruby, on the other hand, line breaks can occur and the layout reverts to the mono ruby layout before and after the line break. The example shows the behavior of our previous 1-3 and 3-1 examples when a line break occurs within the jukugo base character sequence.

note that when there are 3 or more mono ruby characters at the start or end of a line, then the base characters may be moved down to give room, within the Kihonhanmen, for the ruby characters. There are other ways to set the ruby characters in that situation, but that is beyond the scope of this talk.


Concluding remarks
W3C
Concluding remarks
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 49


In conclusion
W3C
• Kihonhanmen provides an alternative way of specifying page layout
• Converting between horizontally and vertically oriented text is not so straightforward
• Tables of character classes provide details on spacial adjustment & line breaking
• There are three forms of ruby: mono, group and jukugo
• The specification is located at: http://www.w3.org/TR/jlreq/
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 50


Thank you
http://www.w3.org/International/
Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 51