Typesetting and layout in multiple directions — Proposed solution

John Plaice

Abstract

I propose a new, general way of looking at typesetting and layout in multiple directions. It subsumes the left-to-right and right-to-left horizontal writing used in most of the world, as well as the vertical writing used in East Asia. The generality allows the development of layout schemes for situations when several writing directions appear on the same page.

The key to the approach is that managing multidirectional text requires a separation of writing style from box direction. It turns out that there are only three different kinds of writing style, and eight kinds of directional box, and that simple rules can be used to define how these different writing styles may appear in the different kinds of box.

1 Introduction

If TEX or a successor thereof is to be used naturally for all of the world's languages, then the problem of typesetting in multiple directions must be solved in all of its generality.

Most of the languages around the world are printed horizontally, from left-to-right, with the first line of the page at the top of the page. In the Middle East, Arabic, Hebrew and several other scripts are written horizontally, from right-to-left. In East Asia, traditional writing and printing is done vertically, with the first line of the page on the right-hand side; in Japan, this practice is still common for literature. Uighur and Mongolian writing is also vertical, but the first line of the page is on the left.

The general problem does not consist in simply printing a text in each of the different directions. A typesetting system to be used for multilingual documents needs to be able to print, on the same page, multiple languages using different direction combinations, as needed. This means that headers, footers, columns, tables, paragraphs, and everything else appearing on a page can potentially appear in the different directions.

In this paper, I propose a general solution that solves not just the above-mentioned cases, but also even more complicated situations, such as for Ancient Greek *boustrophedon* and Rongorongo *reverse boustrophedon*, both with alternating line directions.

The solution, which supposes that the natural writing style of a script is separate from the boxes in which that writing will be embedded, is the result of much reflection examining existing approaches.

2 Previous work

TEX–XETT

In the TEX world, the first work [5] in multidirectionality was made by Donald Knuth and Pierre MacKay, who developed TEX–XETT, which allowed the mixing of left-to-right and right-to-left horizontal texts in the same paragraph. This was done by using nested `\beginL`–`\endL` and `\beginR`–`\endR` pairs to, respectively, embed left-to-right and right-to-left texts. However, their work supposes that all pages are left-to-right, so it is not suitable for true right-to-left documents.

pTEX

Mixed horizontal and vertical typesetting was introduced in the TEX world with pTEX [2], a tool developed at ASCII Corporation in Japan. Still used in Japan, pTEX allows a vertical or a horizontal list to be begun with either of the `\yoko` and `\tate` primitives. In `\yoko` mode, horizontal and vertical boxes have the same meaning as in standard TEX. In `\tate` mode, `\hbox`es are vertical and `\vbox`es are horizontal.

CSS

The work on supporting multiple directions in Cascading Style Sheets (CSS) for HTML [1] has introduced some useful terminology:

Block flow direction: from first to last line;

Inline base direction: from first to last glyph;

Line orientation: direction towards "top" of line.

Omega

Omega was the first successor to TEX to attempt to solve the multidirectional problem in its generality [3, 4, 8]. It assumes that a box or a font's direction can be designated by three characters, where each is one of Top, Bottom, Left, and Right. These characters absolutely designate one of the edges of the physical page. Then a writing direction must designate:

Primary part. The "top" of each page.

Secondary part. The "left" of each line.

Tertiary part. The "top" of each character.

The secondary direction must be orthogonal to the primary direction. The tertiary direction can take all four values. Hence there are 32 possible directions. Here are the most common ones:

TLT — Left-right (LR) scripts, horizontal CJK.

TRT — Right-left (RL) scripts.

RTT — Vertical CJK, upright LR scripts in vertical CJK.

LTL — Mongolian and Uighur (MU).

Vertical CJK: CJK scripts use TLT, LR scripts use RTR, RL scripts use BBR (RBR), and MU scripts use LTR (RTL).

Horizontal: LR and CJK scripts use TLT, and RL and MU scripts “TRT esu” (TRT).

Vertical MU: MU and RL scripts use LTR (LTL), LR scripts use RTR, and CJK scripts use TLT.

Figure 1: Example of Omega text in many directions. The main flow of text is from left to right, but it includes parts that use a number of different directions. Each line explains the general structure for the most common cases of mixing writing directions.

RTL — MU scripts in vertical CJK.
RTR — Rotated LR scripts in vertical CJK.
LTR — Rotated LR scripts in MU.
LTT — Vertical CJK in MU.

Figure 1 demonstrates the use of several of these directions. In the text of a paragraph, a change of `\textdir`, which respects TEX's group nesting, will change the writing direction. In addition, boxes can have a direction definition, as can the page (for headers and footers), the body (of the page), the paragraph, and mathematics.

Notwithstanding the impressive number of possible writing directions, the proposed solution was not sufficiently general, as it did not make provisions for phenomena such as typesetting to a curve. Furthermore, it required different fonts for different writing directions, despite the fact that many of them simply involved rotating text.

3 The solution

In this paper, a completely separate approach, both simpler and more general, is taken: when typesetting, a particular writing style (there are three) is used, and the text will be placed in a secondary (text) box, of which there are eight kinds. Secondary boxes are lined up in a primary box.

The three writing styles are as follows:

- In *baseline left-to-right writing* (`BL`), there is a baseline which the glyphs are sitting upon or hanging from, and successive glyphs are placed to the right of previous ones. Most alphabetic scripts use this form of writing.

  This is baseline left-to-right writing.

- In *baseline right-to-left writing* (`BR`), there is also a baseline, but successive glyphs are placed to the left of previous ones. This form of writing is used for Arabic and Hebrew, but one can also consider the Uighur and Mongolian vertical scripts to be using the same style.

  .gnitirw tfel-ot-thgir esaeline si sihT

- In *axial writing* (`A`), there is an axis flowing through the glyphs, typically in the middle, and the glyphs are pinned onto the axis, one after the other. The vertical typesetting used in Japan, China and Korea is axial writing.

W	T
R	H
I	I
T	S
I	I
N	S
G	A
°	X
	I
	A
	L

The underlying assumption is that a writing system should be invariant to rotation, and not be defined by the box holding it. For example, typesetting along a curve requires no boxes.

TEX's `\vbox`es and `\hbox`es should be, respectively, generalized to *primary* and *secondary* boxes. Here, text will be placed in secondary boxes. Both primary and secondary boxes have a direction and an orientation. The direction is designated by one of `T`, `B`, `L`, or `R`, and the orientation is either positive (`+`) or negative (`-`), where positive or negative refers to the angle made between the direction vector of text and the normal vector pointing towards the part of the

line where annotations such as footnotes are made. Hence, English would normally be placed in a `L+` secondary box, Arabic in a `R-` box, vertical Japanese in a `T+` box, and Mongolian in either a `T+` or `T-` box, depending on where the annotations are placed with respect to the line.

Any piece of text should be embeddable into a text of another style, and into any kind of secondary box. However, when doing so, the text must "adapt" to the box in which it will be placed, possibly by being mirrored or rotated so that it can fit therein.

4 Example embeddings

In this section, we will consider a number of possible embeddings of a text *A* in a text *B*, where the two texts are using different writing styles. In all cases, these texts are:

A: **pneumonoultramicroscopicsilicovolcano-coniosis**

B: I am glad ··· is not well known.

Text *A* was chosen because it is long, with many possible hyphenation points.

Embedding a baseline text in another baseline text

We consider two main cases. In the first, texts *A* and *B* are of the same style. In this case, depending on the fonts, vertical adjustment of the baseline of the embedded text may be necessary.

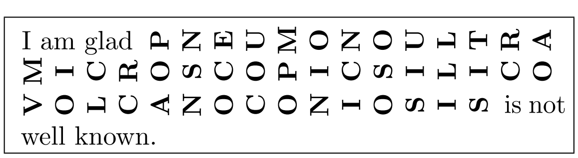
I am glad **pneumonoultramicroscopicsilico-volcanoconiosis** is not well known.

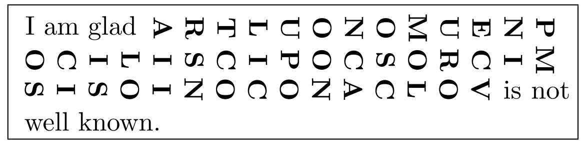
In the second case, texts *A* and *B* are of opposite orientation. There are two subcases. In the first, text *A* is mirrored to resemble the style of text *B*.

In the second subcase, the TeX–XeT bidirectional paragrapher is used to embed text *A*.

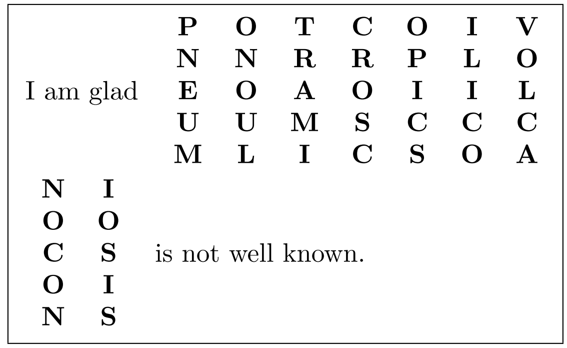
Embedding an axial text in a baseline text

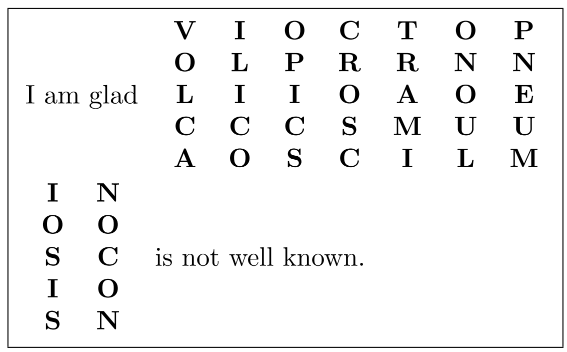
We consider two main cases. In the first, with two subcases, the axial text is rotated, 90° or −90°.





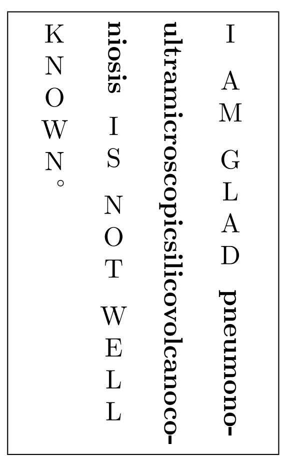
In the second case, the axial text is not rotated, but an inner paragrapher is called upon to create a set of lines perpendicular to the lines of the outer paragraph. For this to work, the length of these inner lines is declared. There are two subcases.

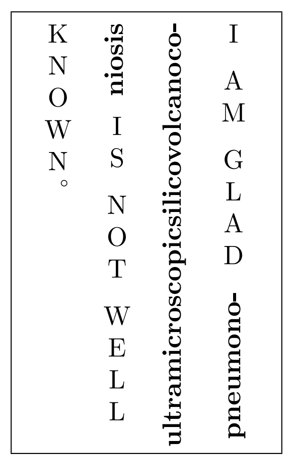




Embedding a baseline text in an axial text

In the first case, the embedded text is rotated.



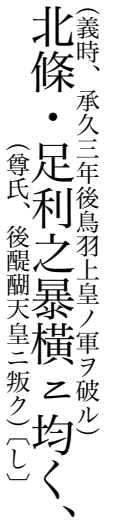


In the second case, the embedded text is not rotated, and an inner paragrapher is called. Once again, the length of the lines of the inner paragraph must be stated.

K	**volcano-**	I
N		A
O	I	M
W	S	
N		G
°	N	L
	O	A
	T	D
	W	**pneumo-**
	E	**noultra-**
	L	**microscop-**
	L	**icsilico-**

5 The importance of box orientation

The direction of a secondary box is not sufficient to completely describe it. Consider the following segment of an example presented by Ken Nakano [7]:



There are two annotation lines, one to the right, one to the left. The box orientation is positive, so the annotation line to the right (resp. to the left) would be considered to be the major (resp. minor) one.

Note that here, there are three parallel streams, each with its own rules for line breaking. This is in fact a special case of multiple interacting streams. See [6] for a more general discussion.

6 Implementation

Adapting existing TEX engines to implement this proposal should be relatively straightforward. There would be a limited number of primitives (text mirroring, clockwise and counterclockwise rotation, vertical and horizontal displacement), using a limited set of parameters: numeric (baseline shift), Boolean (to mirror or not, to rotate or not, clockwise or counterclockwise), and direction (text, mathematics, paragraph, page body, page).

I think that pTEX, XHTEX and LuaTEX can all be extended in these ways, and that much of the relevant code from the Omega project is still useful. The approach is also extendable to new writing styles, e.g., axial bottom-up writing, or to new kinds of text embeddings. The new algorithms to be added include a paragrapher within the paragrapher (straightforward), and a line-breaking algorithm for multiple simultaneous streams (non-trivial).

References

[1] fantasai and Koji Ishii. CSS Writing Modes Level 3. `http://www.w3.org/TR/css3-writing-modes`, W3C Working Draft, 24 October 2013.

[2] Hisato Hamano. Vertical typesetting with TEX. *TUGboat*, 11(3):346–352, September 1990.

[3] Yannis Haralambous and John Plaice. The Omega Typesetting and Document Processing System. *BIT*, 4:137–152, 2001. In Japanese.

[4] Yannis Haralambous and John Plaice. Traitement automatique des langues et composition sous Omega. *Cahiers GUTenberg*, 39–40:139–166, May 2001.

[5] Donald Knuth and Pierre MacKay. Mixing right-to-left texts with left-to-right texts. *TUGboat*, 8(1):14–25, April 1987.

[6] Blanca Mancilla, Jarryd P. Beck, and John Plaice. Typesetting multiple interacting streams. In Cyril Concolato and Patrick Schmitz, editors, *ACM Symposium on Document Engineering*, pages 149–152. ACM, 2012.

[7] Ken Nakano and Hajime Kobayashi. Case study: Typesetting old documents of Japan. *TUGboat*, 34(3):281–284, 2013.

[8] John Plaice and Yannis Haralambous. Writing in multiple directions in Omega. In *Fourth International Symposium on Multilingual Computing*, Tokyo, Japan. 1–3 March 2001.

◇ John Plaice
Montreal, Canada
UNSW, Sydney, Australia
`johnplaice (at) gmail dot com`
`http://plaice.web.cse.unsw.edu.au`