Bidirectional text · Bidirectional text

This page contains some accompanying examples to Alan Flavell's "I18n — text direction". Examples that are supposed to display incorrectly (i.e. not as intended) in either Mozilla or Internet Explorer 6 are in red. Read the source text to understand how it's done!

You can specify text direction by (paired) Unicode control characters, by (paired) control characters written as numeric references, by HTML markup, or by CSS properties. Control characters are restricted to plain text and are not suitable for use with markup languages (except `lrm` and `rlm`). The preferred method for HTML is to use HTML markup. Use control characters written as numeric references only in places where no markup is possible, such as attribute values (`alt`, `title`, etc.). Occasionally it may be convenient to specify text direction via CSS; for example, to set the direction of columns in tables rather than to put a `dir` attribute into each and every `<td>`.

In the following table, `div` represents any block-level element, and `span` represents any inline element.

Plain text control chars	HTML 4 control chars	markup	CSS 2 properties
not applicable	not applicable	`<div dir=ltr>` ..... `</div>`	`direction: ltr;` `unicode-bidi: normal`
not applicable	not applicable	`<div dir=rtl>` ..... `</div>`	`direction: rtl;` `unicode-bidi: normal`
U+202A ..... U+202C	`&#8234;` ..... `&#8236;`	`<span dir=ltr>` ..... `</span>`	`direction: ltr;` `unicode-bidi: embed`
U+202B ..... U+202C	`&#8235;` ..... `&#8236;`	`<span dir=rtl>` ..... `</span>`	`direction: rtl;` `unicode-bidi: embed`
U+202D ..... U+202C	`&#8237;` ..... `&#8236;`	`<bdo dir=ltr>` ..... `</bdo>`	`direction: ltr;` `unicode-bidi: bidi-override`
U+202E ..... U+202C	`&#8238;` ..... `&#8236;`	`<bdo dir=rtl>` ..... `</bdo>`	`direction: rtl;` `unicode-bidi: bidi-override`
U+200E	`&lrm;`	not applicable	not applicable
U+200F	`&rlm;`	not applicable	not applicable

Basic test

If the line below is displayed as "12 11 10 9 8 7 6 5 4 3 2 1 0", then your browser recognizes the `dir` attribute and it is probably ready for right-to-left text. Preferably, the line should be right-aligned.

0 1 2 3 4 5 6 7 8 9 10 11 12

Control (formatting) characters

The control or formatting characters U+202A to U+202E are not suitable for use with HTML. If they are written directly into the source text, they interfere with the left-to-right markup and make editing or even viewing the source a nightmare. Furthermore, the bidirectional algorithm stops at newlines. It would no longer be possible to structure the source text by newlines, which could separate, for example, the paired U+202B and U+202C.

The closing U+202C or `&#8236;` is sometimes implied and may be omitted like the closing `</p>` and `</td>` in HTML. Nevertheless, it is safer to close always explicitly.

To write "תבש [סעבאש]", you can use HTML markup with `<span dir=rtl>` or, exceptionally, write the control characters `&#8235;` and `&#8236;` as numeric references. Inserting the control characters U+202B and U+202C directly results in a mess when viewing the source.

```
&#8235;<B lang="he">■■■</b> [<I>■■■■■</i>]&#8236;

■<B lang="he">■■■</b> [<I>■■■■■</i>]■
```

Advice

Never use UTF-8-encoded control characters, but only character references like `&#8235;` and `&rlm;`.

The `dir` attribute

Three directional levels

Three or more directional levels (here: Latin > Hebrew > Latin) must be defined by control characters or, preferably, by HTML markup. The third line has no `dir` markup and is thus displayed as having only two directional levels.

The words mean “Congratulations!”

The words “מזל טוב” mean “Congratulations!”

The words “מזל [mazel] טוב [tov]” mean “Congratulations!”

The words “מזל [mazel] טוב [tov]” mean “Congratulations!”

Letters and digits

Numbers, which are always written from left to right, are likely to mess with right-to-left text. For example, “12 345” denote two numbers and should be displayed as “345 12”. On the other hand, “12 345” denotes a single number and should always be displayed as “12 345”.

The first line is from Google’s Urdu interface with overall dir=rtl; the second line has proper dir markup. (Both lines are written in the restricted MacUrdu character set.)

© 2004 Google — 90 00 000 ويب صفحات كى تلاش وہ رہى ہے

© 2004 Google — 90 00 000 ويب صفحات كى تلاش وہ رہى ہے

© 2004 Google — 9 000 000 veb safahāt kī talāš ho rahī hai

Advice

Always specify the dir attribute for each piece of text, starting with <body dir=ltr> or <body dir=rtl>.

The **bdo** element

Left-to-right Hebrew

To write Hebrew letters from left to right, you need the bdo element in addition to the attribute dir=ltr.

The vowels α ε η ι ο derive from א ה ח י ע, resp.

The vowels α ε η ι ο derive from א ה ח י ע, resp.

The next examples assume a right-to-left context (dir=rtl) such as an Arabic-language page. The date 31 December 1999 is to be shown in all-numeric form: 1999-12-31. The first line in each example is the one where Internet Explorer 6 fails.

European (North African) digits

The ASCII hyphen is a European number separator. Therefore, no special markup should be necessary. However, Internet Explorer 6 needs dir=ltr.

1999-12-31

1999-12-31

Arabic-Indic digits with non-breaking hyphen

The non-breaking hyphen (‑) is another neutral. Therefore, markup with <bdo dir=ltr> is necessary for all browsers.

١٩٩٩‑١٢‑٣١

١٩٩٩‑١٢‑٣١

Arabic-Indic digits with slash

The traditional Arabic date format calls for the slash as separator and the suffix م (mīlād = birth), meaning “AD”. The slash is a common number separator. Therefore, no special markup should be necessary. However, Internet Explorer 6 needs <bdo dir=ltr>.

١٩٩٩/١٢/٣١ م

١٩٩٩/١٢/٣١ م

Advice

Use the attribute `dir=ltr` with European digits and the tag `<bdo dir=ltr>` with Arabic-Indic digits.

The **lrm** and **rlm** characters

The left-to-right mark (`&lrm; = &#8206;`) and the right-to-left mark (`&rlm; = &#8207;`) are alternative ways to specify the direction of neutral characters such as punctuation marks or spaces. The above examples are rewritten here using `&lrm;`.

Left-to-right Hebrew

The vowels α ε η ι ο derive from א ה ח י ע, resp.

The vowels α ε η ι ο derive from א ה ח י ע, resp.

European (North African) digits

1999-12-31

1999-12-31

Arabic-Indic digits with non-breaking hyphen

١٩٩٩-١٢-٣١

١٩٩٩-١٢-٣١

Arabic-Indic digits with slash

١٩٩٩/١٢/٣١ م

١٩٩٩/١٢/٣١ م

Letters and digits

© 2004 Google — 90 00 000 ويب صفحات كى تلاش وه رهى هے

© 2004 Google — 90 00 000 ويب صفحات كى تلاش وه رهى هے

© 2004 Google — 9000000 ويب صفحات كى تلاش وه رهى هے

The second line does not work in Internet Explorer 5, which needs a number without spaces. This example shows that the explicit markup with the `dir` attribute is more reliable than the implicit `&lrm;` and `&rlm;` marks.

The **zwnj** character

The zero-width non-joiner (`&zwnj; = &#8204;`) is necessary for writing Persian where certain affixes and compound words do not join. It is shown by a hyphen in the transliterated words below.

Persian plurals

هفته	hafteh	week
هفته‌ها	hafteh-hā	weeks
هفته‌ها	haftehhā	wrong
موزه	mūzeh	museum
موزه‌ها	mūzeh-hā	museums
موزه‌ها	mūzehhā	wrong

Compound words

سه	seh	three
سه‌شنبه	seh-šanbeh	Tuesday
سه‌شنبه	sehšanbeh	wrong

راه	rāh	way, road
راه‌آهن	rāh-āhan	railway
راه‌آهن	rāh’āhan	wrong
نرم	narm	soft
نرم‌افزار	narm-afzār	software
نرم‌افزار	narmāfzār	wrong

The **zwj** character

The zero-width joiner (`&zwj;` = `&#8205;`) is necessary to show isolated glyphs of the Arabic letters. At least Mozilla needs it when Arabic letters are separated by HTML markup. (The zero-width joiner does not work with earlier browser versions such as Netscape 7.0 or Internet Explorer 5.)

Markup inside Arabic text

جسيم	jasīm	gros
جسام	jisām	gros pl.
جسيمة	jasīmah	grosse
جسيمات	jasīmāt	grosses
أجسم	ajsam	plus gros(se(s))
الأجسم	al-ajsam	le plus gros
الأجاسم	al-ajāsim	les plus gros
الجسمى	al-jusmā	la plus grosse
الجسمايات	al-jusmayāt	les plus grosses

Isolated glyphs

ن · س · ت · ع · ل · ي · ق ← ن س ت ع ل ي ق ← نستعليق

ن · س · ت · ع · ل · ي · ق ← ن س ت ع ل ي ق ← نستعليق

On the other hand, Internet Explorer 6 joins letters even when they are separated by markup. Therefore you still need an additional `&zwnj;` if the letters shall not join.

سەەزار ، ەەزار

سەەزار ، ەەزار

Urdu aspiration

The zero-width joiner can also be used to write Urdu text in and for the restricted MacUrdu character set where the two-eyed he (`&#1726;`) is not available.

هفته	haftah	week
هاته	hāth	wrong
هاته	hāth	hand
ديده	dīdah	eye
دوده	dūdh	wrong
دوده	dūdh	milk

Sindhi non-connecting he

The sequence `&zwj;&zwnj;` is needed for Sindhi where the initial form of the letter he (ھ) is used as consonant, while the connecting form (ہ) is reserved for aspiration.

جھنگل	jhangalu	jungle
گھر	gharu	house
منھن	munhun	wrong
منہن	munhun	mouth
وڃھ	vīha	wrong
وڃہ	vīha	twenty

Further reading

Persian word processing / ZWNJ — ZWJ

---

xx Andreas Prilop

30 August 2007