

Bidirectional text · Bidirectional text

This page contains some accompanying examples to Alan Flavell's "[l18n — 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.

```

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 ‬ 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 “שבת” (Shabbat), you can use HTML markup with or, exceptionally, write the control characters ‫ and ‬ 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">שבת</b> [<D>שבת</i>]&#8236;
```

```
B> lang="he">שבת</b> [<D>שבת</i>]
```

Advice

Never use UTF-8-encoded control characters, but only [character references](#) like ‫ and ‏.

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 “[tov] טוב [mazel] מזל” 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 ویب صفحات کی تلاش ہو رہی ہے

© 2004 Google — 90 00 000 ویب صفحات کی تلاش ہو رہی 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 ו, resp.

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

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 &lrn; and ‏ marks.

The zwnj character

The zero-width non-joiner (‌ = ‌) 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

sehshanbeh

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 (‍ = ‍) 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

jasīm	م.ي.جسم
gros	
jisām	م.ا.جسم
gros <i>pl.</i>	
jasīmah	ة.م.ي.جسم
grosse	
jasīmāt	ا.ت.م.ي.جسم
grosses	
ajsam	جسم ا
plus gros(se(s))	
al-ajsam	جسم الا
le plus gros	
al-ajāsīm	سم.ا.ج.ا.ل
les plus gros	
al-jusmā	ي.جسم.ا.ل
la plus grosse	
al-jusmayāt	ي.ا.ت.جسم.ا.ل
les plus grosses	

Isolated glyphs

ق.ي.ل.ع.ت.س.ن ← ق.ي.ل.ع.ت.س.ن
ت.ي.ل.ع.ت.س.ن ← ق.ي.ل.ع.ت.س.ن

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

ه.ز.ا.ر.د.ه , ه.ز.ا.ر.س.ه
ه.ز.ا.ر.د.ه , ه.ز.ا.ر.س.ه

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](#) (ھ) is not available.

هفتہ
haftah
week

هاٺ
hāth
wrong

هاٺو
hāth
hand

ڏيڏه
dīdah
eye

دوڏه
dūdḥ
wrong

دودھ
dūdḥ
milk

Sindhi non-connecting he

The sequence ‍‍ is needed for Sindhi where the initial form of the [letter he](#) is used as consonant, while the (ه) connecting form (ه) is reserved for aspiration.

جھنگل
jhangalu
jungle

گھر
gharu
house

منهن
munhun
wrong

منهن
munhun
mouth

ويه
viha
wrong

ويھ
viha
twenty

Further reading

[Persian word processing](#) / [ZWNJ](#) — [ZWJ](#)