# Title

stata.com

**ustrword( ) —** Obtain Unicode word from Unicode string



## Description

`ustrword(`*s*`,` *n*`)` returns the *n*th Unicode word in the Unicode string *s*. Positive numbers count Unicode words from the beginning of *s*, and negative numbers count Unicode words from the end of *s*. `1` is the first word in *s*, and `-1` is the last Unicode word in *s*. The function uses the `locale_functions` setting.

`ustrword(`*s*`,` *n*`,` *loc*`)` returns the *n*th Unicode word in the Unicode string *s*. Positive numbers count Unicode words from the beginning of *s*, and negative numbers count Unicode words from the end of *s*. `1` is the first word in *s*, and `-1` is the last Unicode word in *s*. The function uses the locale specified in *loc*.

`ustrwordcount(`*s*`)` returns the number of nonempty Unicode words in the Unicode string *s*. An empty Unicode word is a Unicode word consisting of only Unicode whitespace characters. The function uses the `locale_functions` setting.

`ustrwordcount(`*s*`,` *loc*`)` returns the number of nonempty Unicode words in the Unicode string *s*. An empty Unicode word is a Unicode word consisting of only Unicode whitespace characters. The function uses the locale specified in *loc*.

When *s* and *n* are not scalar, these functions return element-by-element results.

## Syntax

*string matrix* `ustrword(`*string matrix s*`,` *real matrix n*`)`

*string matrix* `ustrword(`*string matrix s*`,` *real matrix n*`,` *string scalar loc*`)`

*real matrix* `ustrwordcount(`*string matrix s*`)`

*real matrix* `ustrwordcount(`*string matrix s*`,` *string scalar loc*`)`

## Remarks and examples

stata.com

A Unicode word is different from a word produced by the function `word()`. The word in `word()` is a space-separated token. A Unicode word is a language unit based on either a set of word boundary rules or dictionaries for some language such as Chinese, Japanese, and Thai.

An invalid UTF-8 sequence is replaced with a Unicode replacement character `\ufffd`.

The null terminator `char(0)` is a nonempty Unicode word.

## Conformability

`ustrword(`*s*`,` *n*`)`, `ustrword(`*s*`,` *n*`,` *loc*`)`:

| | |
|---|---|
| *s*: | $r \times c$ |
| *n*: | $r \times c$ or $1 \times 1$ |
| *loc*: | $1 \times 1$ |
| *result*: | $r \times c$ |

`ustrwordcount(`*s*`)`, `ustrwordcount(`*s*`,` *loc*`)`:

| | |
|---|---|
| *s*: | $r \times c$ |
| *loc*: | $1 \times 1$ |
| *result*: | $r \times c$ |

## Diagnostics

`ustrword()` returns an empty string if an error occurs. `ustrwordcount()` returns a negative number if an error occurs.

## Also see

[M-5] **invtokens( )** — Concatenate string rowvector into string scalar

[M-5] **tokenget( )** — Advanced parsing

[M-5] **tokens( )** — Obtain tokens from string

[M-4] **string** — String manipulation functions

[FN] **String functions**

**[U] 12.4.2 Handling Unicode strings**