NAME

correl — measure the numerical similarity between two spines

SYNOPSIS

correl [**-f** *templatefile*] [**-m**] [**-p** *n*] [**-s** *regexp*] [*inputfile ...*] [> *outputfile.cor*]

DESCRIPTION

The **correl** command measures the degree of parametric (numerical) similarity between corresponding values in two Humdrum spines. More precisely, **correl** calculates Pearson's coefficient of correlation for paired tokens containing numerical data.

Two modes of operation are provided. In the *single input mode*, a single file containing two equal-length spines is processed. In this mode, the output from **correl** consists of a single number indicating the linear correlation between the two spines of numerical data.

In the *dual input mode*, two single-spine numerical inputs — called the *template file* and the *input file* — are specified by the user. Normally, the template file is considerably shorter than the input file. In this mode, the output consists of a spine of numerical information (`**correl`) that reflects the momentary similarity between the template and the input for each successive moment in the input. In short, the input file is 'scanned' using the template values, and the correlational similarity determined at each point.

In both *single input* and *dual input* modes, output numerical values range between +1 and -1. Correlation values reflect the degree to which two sets of numerical values rise and fall in synchrony. The maximum output value is +1 — indicating that the two sets of numbers are perfectly related according to a linear relationship. A minimum output value of –1 indicates that the two sets of numbers are perfectly out-of-phase — one set of numbers rises while the other set falls by a proportional magnitude, and vice versa. A correlation value of zero indicates that there is no linear relationship between the two sets of numbers.

In *single input mode*, inputs to **correl** must consist of precisely two spines; otherwise an error message is generated and the command is terminated. The two spines may contain different interpretations and represent different types of information. In the case of the *dual input mode*, the input file and template file must have precisely one spine each; the spines may differ in length, but the *template* file must not be longer than the *input* file.

Only numerical signifiers are considered by **correl;** non-numeric input data are ignored. Where a data token contains a mix of numeric and non-numeric signifiers, only the first complete numerical subtoken contributes to the calculation. The following examples illustrate how **correl** interprets mixed data tokens:

data token	numerical interpretation
`4gg#`	4
`4.gg#`	4
`-33aa`	-33
`-aa33`	33
`x7.2yz`	7.2
`a7..2bc`	7
`[+5]12`	5
`$17@2`	17
`a1b2 c.3.d`	1 0.3

Humdrum multiple-stops require special attention in **correl** (see below).

In the *dual input mode* the output from the **correl** command consists of a set of records matching the structure of the *input* document. Output values indicate the correlation between the *template* data values and the input data values beginning at that record.

When the *dual input mode* is invoked, it is recommended that output files produced by the **correl** command should be given names with the distinguishing '.cor' extension.

OPTIONS

The **correl** command provides the following options:

-f *templatefile*	specify source pattern as *templatefile* and invoke dual input mode
-h	displays a help screen summarizing the command syntax
-m	disable matched-pairs criterion
-p *n*	output precision to *n* decimal places
-s *regexp*	skip; completely ignore data records matching *regexp*

Options are specified in the command line.

The **-f** option is used to specify an independent *template file* and so invoke the *dual input mode*.

The **-p** option can be used to set the precision of the output values to *n* decimal places. The default precision is 3 decimal places.

The **-s** option allows the user to avoid (or skip) the processing of certain types of data records. This option must be accompanied by a user-defined regular-expression. Input data records matching this expression are not processed.

Correlation values can be calculated only where all numerical data are arranged as matched pairs — that is, the input conforms to the "matched pairs criterion." For example, the following two spines illustrate numerical data matching. The number of numerical data values in both spines are matched throughout the inputs:

```
**spine1    **spine2
10.0        4
7 3         2 .91
.           .
13.8        4
5 8 5       1 1 2
a b c       x
.           p q
*-          *-
```

By contrast, the following file shows several transgressions of the matched pairs criterion. For example, the first data record gives a numerical value in spine #1 that is not matched by a numerical value in spine #2. Similarly, the multiple-stop values in the second data record are unmatched in spine #2:

```
**spine1    **spine2
9.7         a
7 31        2
.           114
426         .
r 11 7      35 xy08z 28
a b c       6 .07
.           p q
*-          *-
```

In normal operation, a single failure to conform to the matched pairs criterion will cause **correl** to issue an error message and terminate operation. If the **-m** option is invoked, unmatched data is simply ignored. For example, with the **-m** option, the above input is treated as equivalent to the following input:

```
**spine1    **spine2
.           .
7           2
.           .
.           .
11 7        35 08
.           .
.           .
*-          *-
```

EXAMPLES

The following examples illustrate the operation of **correl.** The first example shows an excerpt containing considerable parallel motion between two polyphonic voices. Measuring the pitch-contour similarity can be done using the single input mode.

```
!! J.S. Bach, Invention No. 8; BWV 779
**semits        **semits
*M3/4           *M3/4
9               17
12              21
10              19
12              21
9               17
12              21
10              19
12              21
=6              =6
5               14
9               17
7               16
9               17
*-              *-
```

In order to avoid processing the measure numbers, the skip (**-s**) option is used; executing the command:

```
correl -s = bwv779
```

will produce the following output:

```
0.979
```

The second example illustrates the dual input mode. The target input consists of a single spine (labelled `**input`) containing mixed alphabetic and numerical values. (This input file is shown below as the left-most spine.) The template file consists of the numerical sequence: 1, 2, 3 — mixed with the letters a, b, c. (This file is shown as the middle spine below.) Note that the non-numeric characters in both the input and template files have no influence on the operation of **correl.** The third (output) spine is produced by the following command:

```
correl -f template input > output.cor
```

```
(input      (template     (correl
 file)        file)       output)
**input     **template    **correl
0           1a            1.000
1           2b            1.000
2           3c            1.000
3           *-            -0.655
4                         -0.655
x1x                       0.866
y2.                       0.866
2z                        0.000
(3)                       -1.000
[2]                       .
01                        .
*-                        *-
```

The similarity values generated by **correl** are given in the `**correl` spine. Each successive value in the output spine is matched with a data token in the target input file (`**foo`). For example, the initial three output values (1.000) indicate that exact positive correlations occur between the template and the input. That is (0, 1, 2) (1, 2, 3) and (2, 3, 4) all show simple equi-distant increases corresponding to the source template. The final numerical value in `**correl` shows a negative correlation (-1.000) indicating that the numerical sequence (3, 2, 1) is the exact opposite contour to the source template (1, 2, 3). By contrast, the immediately preceding output value (0.000) indicates that the sequence (2, 3, 2) shows no systematic linear relationship with the source template (1, 2, 3).

The following example provides a more complicated illustration of **correl.** Once again the left-most spine is the target input, the middle spine is the source template, and the right-most spine shows the corresponding output.

```
(input      (template     (correl
 file)        file)       output)
**input     **template    **correl
=1          1             .
1           2 3           1.000
2 3         .             -0.370
100         4             -0.742
8r          5 6           .
4           *-            0.042
5 6                       .
=2                        .
0                         .
4r                        .
-2x -3                    .
-x8                       .
==                        .
*-                        *-
```

The above output spine was created by executing the command:

```
correl -m -s '[=r]' -f template input > output.cor
```

Due to the **-s** option, all records in the input file containing an equals-sign or lower-case 'r' are eliminated from the calculations. The presence of the null-token in the third data record of the template file is noteworthy. Although no correlations are calculated with the null-token, it acts as a place-holder, and causes the corresponding record in the input file to be ignored. For example, the first correlation value is calculated on the basis of the following coordination of numerical data:

```
1       1
2 3     2 3
100     .
4       4
5 6     5 6
```

Since the value '100' is not matched with a numerical value in the template, it is ignored in the correlation measure. (Note that without the **-m** option, no output would be generated.)

At the next instant, the correlation value is calculated on the basis of the following coordination of numerical data:

```
2 3     1
100     2 3
4       .
5 6     4
0       5 6
```

The double-stops do not form matched pairs, hence much of the data is discarded. For example, in the first data record, 2 is matched with 1 and 3 is discarded. In the second record, 100 is matched with 2 and 3 is discarded, etc.

The third correlation value is calculated on the basis of the following coordination of numerical data:

```
100       1
4         2 3
5 6       .
0         4
-2 -3     5 6
```

In this case, the correlation value is based on the following numerical pairing: 100 ⇔ 1, 4 ⇔ 2, 0 ⇔ 4, –2 ⇔ 5, –3 ⇔ 6. All other numerical values are ignored.

The final correlation value in this example is calculated on the basis of the following coordination of numerical data:

```
4          1
5 6        2 3
0          .
-2 -3      4
8          5 6
```

The corresponding correlation value is based on the following numerical pairing: 4 ⇔ 1, 5 ⇔ 2, 6 ⇔ 3, –2 ⇔ 4, 8 ⇔ 5.

PORTABILITY

DOS 2.0 and up, with the MKS Toolkit. OS/2 with the MKS Toolkit. UNIX systems supporting the *Korn* shell or *Bourne* shell command interpreters, and revised *awk* (1985).

SEE ALSO

patt (4), **pattern** (4), **simil** (4)

WARNINGS

Correlation coefficients indicate only the magnitude of the association between two sets of data. High correlation values can occur purely by chance. The noteworthiness (statistical significance) of a correlation value depends on the number of input values given in the template spine. Novice users should consult a standard statistics textbook for further advice on how to interpret the results.

For formal statistical measures, the **-m** option should never be invoked.

If only one pair of matched values is present, the linear correlation is mathematically undefined. In this case a question mark signifier is output.

LIMITS

The **correl** command is currently unable to handle input files greater than about 4,000 records.