DATA Step

I have a list of names with initials some of which have a period but not all - can I add a period after each initial?

Yes! This is one of those data cleansing routines to create consistent data.

The process involves SCANning the character string extracting each word in turn, then if it has a length of 1 replace it with a new string consisting of the initial letter with a period concatenated to it.

It is possible that there are some one-letter 'words' in the string which are not initials, e.g. an ampersand '&' so remember to exclude these from the logic.

Firstly, let's create some dirty data on which to operate:

```
data names ;
  infile datalines dsd ;
  input name1 :$40. ;
datalines ;
Mr N. D. Beale
Mrs P P J Smith
Mr & Mrs D Williams
Baron V von Hohenzollern
Miss G Green
Mr J & Mrs N Chambers
;
run ;
```

There are varying numbers of words within each name, and some have periods after the initials, but others do not. The first thing to ascertain is the maximum number of words in the longest name, and we will use this as the upper bound of a DO-END loop. Creating a macro variable to contain the maximum word count:

```
proc sql noprint ;
  select max(countw(name1)) into :maxw separated by ''
  from names
  ;
quit ;
```

Then we can start to process each name in turn.

```
data names (keep = name1) ;
  set names ;
  length word worddot $ 2 ;
  do i = 1 to &maxw until (word = '') ;
    word = scan(name1,i,' ') ;
    call scan(name1,i,posn,len,' ') ;
    if len = 1 and word not in ('&' '') then
    do ;
      worddot = cats(word,'.') ;
      name1 = catx( ' '
                   ,substr(name1,1,posn-1)
                   ,worddot
                   ,substr(name1,posn+len)
                  ) ;
    end ;
  end ;
run ;
```

The conditional DO-END loop reads each i'th word in turn, up to the maximum number of words, or the returned word is NULL. Within the loop the SCAN function extracts the i'th word, while the CALL SCAN routine returns the position of the first character of the word, and its length.

We now have the three elements:
word
posn
len
and can start to build and replace the single-character initials.

The conditional logic determines whether the i'th word has a length of 1, and is neither an ampersand, nor NULL. We create the new element worddot by adding a period to the initial. We then rebuild the name string by concatenating the text up to the position of the i'th word (posn - 1), the new element (worddot), and the remainder of the string after the i'th word (posn + len).

Unique solution ID: #1074
Author: Alan D Rudland
Last update: 2023-08-23 11:42