

Creation of *Fontcase* Macro in SAS® to Generate Texts of Desired Capitalization

Ranjan Sitaula, PhD, The EMMES Company, LLC

ABSTRACT

Standard clinical datasets often contain character variables with texts. Some of the examples of such variables include adverse events terms, disposition reason and comment fields. There is often a requirement to have prespecified capitalization on these variables. Moreover, the raw datasets, which are the primary source for these datasets often contains text strings with typographical errors and random capitalization. Use of standard SAS functions such as *propcase* or *lowcase* do not always give the desired capitalization. Similarly, use of characters replacement functions such as *tranwrd* and *translate* functions to fix capitalization and typographical errors can be cumbersome when multiple words are involved. To address this shortcoming, a single macro program called “Fontcase Macro” has been created which performs the following tasks simultaneously with minimal user input: (i) converts texts of any prior capitalization to sentence case or title case as required, (ii) converts or maintains abbreviations/standard terminologies to standard capitalization, (iii) corrects typos. The macro code uses the Perl Regular Expression (RegEX) function *prxchange* to identify text patterns and change them to desired capitalization. Moreover, prespecified abbreviations and standard terminologies are stored in a spreadsheet file that is accessed by the macro program. The macro has been extremely useful in the generation of standard clinical datasets such as SDTM (Study Data Tabulation Model) and ADaM (Analysis Data Model) datasets for variables involving text strings and in the generation of TLFs (tables, listings, and figures) using raw datasets.

INTRODUCTION

SAS programmers very frequently come across instances where we are dealing with character variables containing lengthy texts and strings. If these variables are used in the generation of output tables or standard datasets, it is often required to alter the capitalization of these text strings. In clinical trials study, standard datasets such as SDTM and ADaM datasets are generated using raw data which contain inconsistent capitalization and typos. And trying to find a correct capitalization, or in other instances, dealing with typos is a frequent occurrence and is often time consuming. A program that addresses for these capitalization requirements and addresses the inaccuracy issues with minimal user input saves significant programming time. The paper presents a macro program, *fontcase* macro, which was created to overcome these issues. The macro incorporates the following tasks:

1. Convert texts of any capitalization to either sentence case or title case as desired.
2. Adjust standard terms and abbreviations to prespecified capitalization.
3. Correct typos.

Hence, the macro can be for the sole purpose of correcting typos and/or abbreviations, the sole purpose changing capitalization, or a combination of both. It also offers the flexibility to take variable values of any capitalization and change them to sentence case or title case.

The macro primarily uses the SAS functions *propcase* and *lowcase* in conjunction with the powerful Perl Regular Expression (Regex) to get the desired capitalization and correct typos and

abbreviations/standard terms. While the pattern matching and replacement ability of Regex helps in the replacement of incorrect terms, the pattern matching and replacement ability together with positive and negative lookbehind and lookahead features of Regex help in assigning correct capitalization by identifying standalone words and the beginning and end of sentences. In addition to the program code, the macro reads from a separate spreadsheet containing standard and abbreviated terms with fixed capitalization and uses the *prxchange* Regex function to find and replace the texts in character variables. The use of a separate spreadsheet ensures minimal user intervention in the macro program code and provides flexibility to the user to make required additions/edits to the list of terms in the spreadsheet on a need basis.

MACRO PARAMETERS AND DEFINITION

The following is the macro with parameters used in fontcase macro:

```
%fontcase (data =, fcase =, varlist =, oldval =, newval =, listpath =);
```

where *data* is the SAS dataset of interest,
fcase is the desired capitalization. *fcase* is assigned a value of *propcase* for title case, *lowercase* for sentence case, and left blank for no change in existing capitalization,
varlist is the variable in the dataset you want to run the macro on,
oldval is the word/string that you want to search and replace. *oldval* can be multiple words separated by comma,
newval is the corresponding replacement value(s) for *oldval*,
Listpath is the filepath and filename of the spreadsheet containing terms with prespecified capitalization.

Among the macro variables, *fcase*, *oldval*, *newval* and *listpath* can assume blank values depending on existing variable values. A blank value for *fcase* denotes that no change in capitalization is required. Blank values for *oldval* and *newval* denote that there are no typos in the variable filed that needs correction, and a blank value for *listpath* denote that no correction is required for standard and abbreviated terminologies.

Figure 1 shows the flow diagram for the execution of fontcase macro. The macro is run with the initial set of assignments for the macro parameters. If the output variable requires further correction after the initial run, then we update the macro parameter accordingly. In most cases, the updates involve adding term(s) in the spreadsheet and updating the terms specified in the *oldval* and *newval* macro parameters.

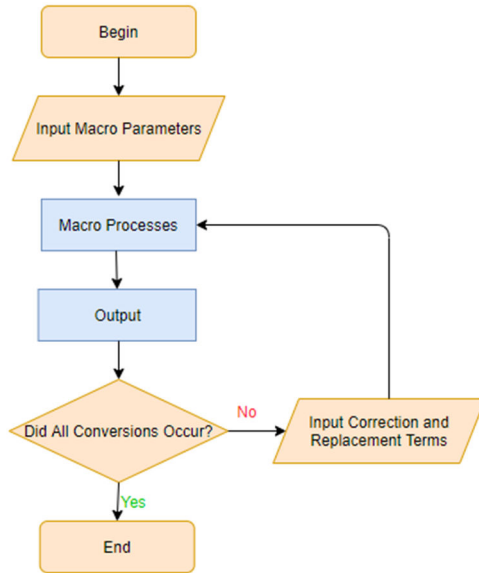


Fig 1. Flow diagram of fontcase macro execution

Figure 2 shows the snippet of the spreadsheet that contains the standard abbreviation terms. The terms in the spreadsheet is a custom list for the variable values used in this paper. The spreadsheet is unique to our dataset and can be updated anytime during the program run to accommodate for any missing terms.

	A	B
1	WBC	
2	EDTA	
3	HGB	
4	ALT	
5	GGT	
6	ALP	
7	LIP	
8	AMY	
9	AST	
10		

Fig 2. Snippet of excel spreadsheet containing abbreviated and standard terminologies.

PROGRAMMING CODE AND STEPS

The following is the sequential order of programming steps and code:

Step1: Locate input dataset and data library, if reading from external location.

```
libname input "C:\temp";
```

Step2: Identify lowercase terms and assign as a macro variable.

```
*****
```

Assign macro variable "lcase" to the terms that are always lowercase even for Title case separated by "|". Macro variable list contains the following parts of speech.

- articles
- Short prepositions (3 letter or less)
- Short conjunctions (3 letter or less).

*****;

```
%let lcase =  
%str(a|an|the|as|at|but|by|for|in|of|off|on|per|to|up|via|and|but|for|nor|or|  
so|yet);
```

Step 3: Import standard and abbreviation terms from spreadsheet and assign as macro variables.

Import standard and abbreviation terms from the Excel spreadsheet. Path and filename are specified in the macro parameter "listpath". The spreadsheet contains a single column and lists individual terms in desired capitalization in individual rows. The program code looks for the terms on the list irrespective of case sensitivity but replaces with the capitalization as specified in the spreadsheet.

*****;

```
proc import datafile="&listpath." out=temp1 dbms=excel replace;  
    getnames=no;  
run;
```

Remove "/" from the spreadsheet terms, if any since it generates syntax error while using the PRXCHANGE statement. The "/" term will be reinstated at a later step.

*****;

```
data temp2;  
    set temp1;  
    term=tranwrd(F1, '/', '###');  
run;
```

List all the spreadsheet terms as a macro variable "abbrev" separated by "|".

*****;

```
proc sql noprint;  
    select term into: abbrev separated by "|" from temp2;  
quit;  
%put &abbrev.;
```

Step 4: Read dataset from input library and perform manipulations for desired capitalization.

```
data &data.;  
    set input.&data.;  
    options noquotelenmax;  
    %let count1 = %sysfunc(countw(%quote(&abbrev), %str(|) ));  
    %let count2 = %sysfunc(countw(%quote(&oldval), %str(,) ));
```

```

%let var1=&varlist.;

*****
Change all the variable texts to "propcase" or "lowcase", or leave as is
depending on the output requirement.
*****;

%if &fcase. = lowercase or &fcase.=propcase %then %do;
    &var1. = &fcase.(&var1.);
%end;

*****
Change abbreviations/standard terms to prespecified capitalization as listed
on the spreadsheet. Reinstate the "/" term if it was present and was earlier
converted to "###"
*****;

&var1.= tranwrd(&var1., '/', '###');
%do m = 1 %to &count1.;
    %let old4 = %scan(&abbrev, &m, "|");
    &var1. = prxchange("s/\b&old4.\b/&old4./i", -1, &var1.);
%end;

&var1. = tranwrd(&var1., '###', '/' );

*****
Change articles, short prepositions, and short conjunctions stored in the
macro variable "lcase" to lowercase for title case output. Code uses regex
negative lookbehind and negative lookahead to ensure that only standalone
terms are replaced. For example, it identifies and changes "or" but not
"Organ" during title case conversion.
*****;

%if &fcase. = propcase or &fcase.=lowcase %then %do;
    &var1. = prxchange("s/(?<!\w|\||\') (&lcase.) (?!\w|\/)/\L$1\E/i",
-1, &var1.);
%end;

*****
Capitalize beginning word of a sentence to uppercase for title case and
sentence case output. Code used uses regex positive lookbehind.
*****;

%if &fcase.= lowercase or &fcase.=propcase %then %do;
    &var1. =
    prxchange('s/((^\s*\w)|(?<=\.|?|\!|\:)(\s*\w))/\U$1\E/',-
1,&var1);
%end;

*****
Capitalize end word of a sentence to uppercase for Title case output.
*****;

%if &fcase.= propcase %then %do;
    &var1. = prxchange("s/(\w+)(?=\.|?|\!|\:)/\u\L$1\E/", -1, &var1);
%end;

```

```
*****
Change terms containing numbers and characters, such as Subject IDs, dates,
temperature values etc., to uppercase. Eg, subject ID of 01XYZ2345 or a date
of 01JAN2024.
*****;
```

```
&var1. = prxchange("s/((\d+\-\w+\d?)|(-?\d+\w?)|(\d+\-\w+)|(\w+\d+))/\U$1\E/i",-1,&var1.);
```

```
*****
Correct Typos by replacing terms assigned to the macro variable "oldval" to
terms assigned to the macro variable "newval"
*****;
```

```
%do n=1 %to &count2.;
    %let old5 = %scan(&oldval., &n, ',');
    %let new5 = %scan(&newval., &n, ',');
    &var1. = prxchange("s/\b&old5.\b/&new5./i",-1, &var1.);
%end;
```

```
run;
```

```
%mend fontcase;
```

Step 5: Macro Call.

```
*****
(a) Example of macro call to change CHARVAR variable values to sentence case.
In addition to changing capitalization, the terms "Bili" and "creat" are
corrected to "bilirubin" and "creatinine" by assigning parameter values in
oldval and newval.
*****;
```

```
%fontcase (data=test,fcase=lowcase, varlist=%str(CHARVAR),
    oldval=%str(bili, creat), newval=%str(bilirubin, creatinine),
    listpath=%str(&filedir.\abbrevlist.xlsx));
```

```
*****
(b) Example of macro call to change CHARVAR variable values to title case.
The terms "Bili" and "creat" are corrected to "Bilirubin" and "Creatinine" by
assigning parameter values in oldval and newval.
*****;
```

```
fontcase (data=test,fcase=propcase, varlist=%str(CHARVAR),
    oldval=%str(bili, creat), newval=%str(Bilirubin, Creatinine),
    listpath=%str(&filedir.\abbrevlist.xlsx));
```

```
*****
(c) Example of macro call where capitalization of CHARVAR variable values is
left unchanged. The only correction here is the capitalization on standard
abbreviation terms as listed in the "abbrevlist" spreadsheet.
*****;
```

```
%fontcase (data=test,fcase=, varlist=%str(CHARVAR),
    oldval=%str(), newval=%str(),
    listpath=%str(&filedir.\abbrevlist.xlsx));
```

OUTPUT

Figure 3 shows outputs of the conversion from a test dataset containing character variable CHARVAR. Section *a* is the variable in its original form, As seen in the figure, the variable texts are inconsistent in terms of capitalization. Among the highlighted terms, “hgb” and “wbc” should be uppercase since these are standard abbreviated medical terms. Similarly, we want to replace the terms “creat” and “Bili” with “Creatinine” and “Bilirubin”. Referring to the macro call in Step 5, the abbreviations are corrected using the “abbrevlist.xlsx” spreadsheet (Fig 2), and the replacement terms are updated using the *oldval* and *newval* macro parameters. Section *b* in figure 3 shows the conversion to sentence case and section *c* shows the conversion to title case. In addition to changing the texts to desired capitalization, the terms mentioned earlier are corrected as highlighted in the figure sections. In the output shown in section *d*, the capitalization is left unchanged, and the only updates are to the abbreviated terms that are capitalized. The output in the figure shows versatility and flexibility offered by the macro.

SN	CHARVAR	SN	CHARVAR	SN	CHARVAR	SN	CHARVAR
1	Hemoglobin Comment: this is a test record for normal hgb male; Platelets, EDTA Comment: this is a test record for abnormal platelet; WBC Comment: this is a test record for normal wbc ; Creatinine Comment: this is a test record for normal creat male; ALT Comment: this is a test record for abnormal ALT male; Bilirubin Comment: this is a test record for normal Bili male;	1	Hemoglobin comment: This is a test record for normal HGB male; platelets, EDTA comment: This is a test record for abnormal platelet; WBC comment: This is a test record for normal WBC ; creatinine comment: This is a test record for normal creatinine male; ALT comment: This is a test record for abnormal ALT male; bilirubin comment: This is a test record for normal bilirubin male;	1	Hemoglobin Comment: This Is a Test Record for Normal HGB Male; Platelets, EDTA Comment: This Is a Test Record for Abnormal Platelet; WBC Comment: This Is a Test Record for Normal WBC ; Creatinine Comment: This Is a Test Record for Normal creatinine Male; ALT Comment: This Is a Test Record for Abnormal ALT Male; Bilirubin Comment: This Is a Test Record for Normal Bilirubin Male;	1	Hemoglobin Comment: this is a test record for normal HGB male; Platelets, EDTA Comment: this is a test record for abnormal platelet; WBC Comment: this is a test record for normal WBC ; Creatinine Comment: this is a test record for normal creat male; ALT Comment: this is a test record for abnormal ALT male; Bilirubin Comment: this is a test record for normal Bili male;
2	ALT Comment: Normal Female ALT; Bilirubin Comment: Grade 1 w/ no ALT increase;	2	ALT comment: Normal female ALT; bilirubin comment: Grade 1 w/ no ALT increase;	2	ALT Comment: Normal Female ALT; Bilirubin Comment: Grade 1 W/ No ALT Increase;	2	ALT Comment: Normal Female ALT; Bilirubin Comment: Grade 1 w/ no ALT increase;
3	Gamma-Glutamyl Transferase Comment: Screened Ineligible; Alkaline Phosphatase Comment: Screened Ineligible; Serum Lipase Comment: Screened Ineligible; Serum Amylase Comment: Screened Ineligible; Aspartate Amino Transferase Comment: Screened Ineligible;	3	Gamma-glutamyl transferase comment: Screened ineligible; alkaline phosphatase comment: Screened ineligible; serum lipase comment: Screened ineligible; serum amylase comment: Screened ineligible; aspartate amino transferase comment: Screened ineligible;	3	Gamma-Glutamyl Transferase Comment: Screened Ineligible; Alkaline Phosphatase Comment: Screened Ineligible; Serum Lipase Comment: Screened Ineligible; Serum Amylase Comment: Screened Ineligible; Aspartate Amino Transferase Comment: Screened Ineligible;	3	Gamma-Glutamyl Transferase Comment: Screened Ineligible; Alkaline Phosphatase Comment: Screened Ineligible; Serum Lipase Comment: Screened Ineligible; Serum Amylase Comment: Screened Ineligible; Aspartate Amino Transferase Comment: Screened Ineligible;
4	Gamma-Glutamyl Transferase Comment: Screened Ineligible;	4	Gamma-glutamyl transferase comment: Screened ineligible;	4	Gamma-Glutamyl Transferase Comment: Screened Ineligible;	4	Gamma-Glutamyl Transferase Comment: Screened Ineligible;
5	WBC Comment: this is a test record for normal wbc ; Creatinine Comment: this is a test record for normal creat male; ALT Comment: this is a test record for abnormal ALT male; Aspartate Amino Transferase Comment: AST; Bilirubin Comment: this is a test record for normal Bili male; Gamma-Glutamyl Transferase Comment: GGT; Alkaline Phosphatase Comment: ALP; Serum Lipase Comment: LIP; Serum Amylase Comment: AMY;	5	WBC comment: This is a test record for normal WBC ; creatinine comment: This is a test record for normal creatinine male; ALT comment: This is a test record for abnormal ALT male; aspartate amino transferase comment: AST; bilirubin comment: This is a test record for normal bilirubin male; gamma-glutamyl transferase comment: GGT; alkaline phosphatase comment: ALP; serum lipase comment: LIP; serum amylase comment: AMY;	5	WBC Comment: This Is a Test Record for Normal WBC ; Creatinine Comment: This Is a Test Record for Normal creatinine Male; ALT Comment: This Is a Test Record for Abnormal ALT Male; Aspartate Amino Transferase Comment: AST; Bilirubin Comment: This Is a Test Record for Normal Bilirubin Male; Gamma-Glutamyl Transferase Comment: GGT; Alkaline Phosphatase Comment: ALP; Serum Lipase Comment: LIP; Serum Amylase Comment: AMY;	5	WBC Comment: this is a test record for normal wbc ; Creatinine Comment: this is a test record for normal creat male; ALT Comment: this is a test record for abnormal ALT male; Aspartate Amino Transferase Comment: AST; Bilirubin Comment: this is a test record for normal Bili male; Gamma-Glutamyl Transferase Comment: GGT; Alkaline Phosphatase Comment: ALP; Serum Lipase Comment: LIP; Serum Amylase Comment: AMY;
(a)		(b)		(c)		(d)	

Fig 3. Output of conversion of character variable values to different capitalization. (a) is the variable values in original form, (b) is the conversion to sentence case, (c) is the conversion to title case, (d) is the correction of just the abbreviation terms.

LIMITATIONS, PITFALL AND REMEDY

Conversion to sentence case is not recommended if the character strings include lot of proper nouns. The macro does not correct for proper nouns, unless specified in the spreadsheet or “*oldval*”/“*newval*” macro parameters. And the process can be cumbersome if there are a lot of proper names involved.

As for the pitfalls of using this macro, there are chances of few errors being introduced in the variable value because of automatic conversion of all terms to *proprcase* or *lowcase*. Such errors can be remedied using the following steps:

- Enter the erroneous term and correction term in the “oldval” and “newval” parameters and rerun the macro, leaving all the other parameters unchanged in macro invocation.
- Enter correct capitalization for the incorrect terms in the spreadsheet.

CONCLUSION

The paper provided the detailed code of the *fontcase* macro and provided examples of usage for different conditions. As seen from the examples, the macro offers versatility to update character variables depending on the users’ needs. Although an example was not presented in this paper, the macro can also be used equally effectively to change upper case texts to sentence/title case. Working with character variables comes with many challenges in the form of capitalization, typographical and grammatical issues and may often be time consuming. *fontcase* macro can be a very useful addition to tackle such issues.

ACKNOWLEDGEMENTS

The author would like to thank his coworkers at The Emmes Company for their valuable feedback in creating this macro.

REFERENCES

- <https://support.sas.com/resources/papers/proceedings20/5172-2020.pdf>
- <https://www.regular-expressions.info/lookaround.html>
- https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/lefunctionsref/p1vz3ljudbd756n19502acxazevk.htm
- <https://apastyle.apa.org/style-grammar-guidelines/capitalization/title-case>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact the author at:

Ranjan Sitaula
The Emmes Company, LLC
401 N Washington St., # 700
E-mail: rsitaula@emmes.com