### Paper 315- 2009

# New Language Features in SAS® 9.2 for the Global Enterprise Manfred Kiefer, SAS Institute Inc., Heidelberg, Germany

### **ABSTRACT**

Successful global enterprises speak the languages of their customers. To be successful in the global economy, software supporting international business must also "speak" multiple languages. The SAS® 9.2 Unicode Server not only supports storing, processing, and delivering multilingual data, but also allows switching the language of the user interface.

This paper covers language switching and linguistic sorting. It also touches on multilingual computing.

### INTRODUCTION

Being a global company with a global customer base, SAS® has been translating software and documentation for a long time. Last year, SAS products were localized (or translated) into 24 different languages or language variants. Our customers are demanding that SAS software be available in all local language markets that we sell to. Moreover, SAS products should work in all of our global markets. An internationalized product will process data in a user's native language and will correctly support their formats for dates, times, numbers, currencies, and more. Multi-language data support is a mandatory requirement for running a global business.

Let us look at a simple sample scenario, since a concrete example usually makes it easier to follow an abstract topic like this.

A company has received name and address source data from several Western European countries and has imported it into their database. The data is generated into enhanced contact details for direct mail communications in SAS.

The company intended to process names and addresses from Eastern European and Asian countries. However, some or all of the characters were corrupted after being imported into SAS because they did not use an appropriate encoding for the particular languages. To avoid this problem, they could change the encoding each time they process a language that belongs to a different group.—for example, by running SAS with ENCODING=WLATIN1 for French data, and then rerunning the script with ENCODING=WLATIN2 for Polish, and with ENCODING=SHIFT-JIS for Japanese. The result is multiple output data sets.

However, a better approach is to use an encoding that deals with all language groups (for example, Western European, Eastern European, and Japanese) together and to generate a single output data set. This encoding is UTF-8, a form of Unicode, a universal encoding that can handle characters from all possible languages. With the SAS® Unicode Server, which is a SAS session encoding of UTF-8, it is possible to store, process, and deliver multilingual data.

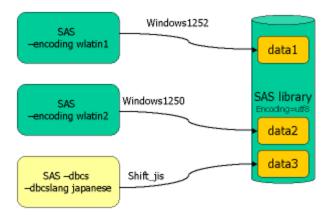


Figure 1: Creating a Multilingual Library

### WHAT'S NEW IN 9.2?

SAS® 9.2 expands the scope and capabilities of National Language Support (NLS). Many features are easier to use and more intuitive. You can not only switch locales and languages "on the fly," but also view and edit SAS program content across multiple languages within your SAS applications.

The SAS program editor (known as the "Enhanced Editor" on Windows) also added Unicode support in SAS 9.2. This means that you can handle multilingual SAS program content within your SAS applications.

Is this a big deal? Let us illustrate the impact by reviewing some before-and-after pictures. Here is a SAS program that features multiple languages as it would appear in SAS 9.1.3 or in SAS<sup>®</sup> Enterprise Guide<sup>®</sup> 4.1:

```
/* summarize the data across a category */

/* Podsumowania danych pochodzacych z róznych kategorii (Polish)*/

/* ???????????????? (Japanese)*/

/* ???????? (Simplified Chinese)*/

/* ?? ??? ?? ???? (Korean)*/

□ proc means data=&data &stat noprint;

var &measure;

class &category &report;

output out=summary &stat=&measure &category /levels;

run;
```

Figure 2: SAS Language Editor in SAS Enterprise Guide 4.1

Notice all of the question marks. That is what your content looks like when characters cannot be rendered in the application's current encoding setting. Now look at the same program as viewed in SAS 9.2 or in SAS Enterprise Guide 4.2:

```
■ Save ・ New ・ Stop Select Server | Export ・ Send To ・ Create ・ Properties

/* summarize the data across a category */
/* Podsumowania danych pochodzących z różnych kategorii (Polish)*/
/* 要約するカテゴリーを越えてデータ (Japanese)*/
/* 总结各地的数据类 (Simplified Chinese)*/
/* 요약 범주에 걸쳐 데이터를 (Korean)*/
□ proc means data=&data &stat noprint;
    var &measure;
    class &category &report;
    output out=summary &stat=&measure &category /levels;
    run;
```

Figure 3: SAS Language Editor in SAS Enterprise Guide 4.2

Does this mean that these multilingual programs will work correctly when you run them in SAS? When using the SAS® Unicode Server, as described above, the answer is yes.

The following information outlines additional features that are new or that have been enhanced in SAS 9.2:

- There are a number of new locales such as Afrikaans\_SouthAfrica, Arabic\_Yemen, Bengali\_India, Catalan Spain, Hindi India, and so on..
- New encodings such as e137 Devanagari EBCDIC have been added.
- The following SAS 9.2 NLS formats and informats have been enhanced and now support Arabic: \$LOGVS, \$LOGVSR, \$VSLOG, and \$VSLOGR.
- The NLBEST format writes the best numerical notation, based on the locale.
- A number of currency formats and informats are new for SAS 9.2. These formats write or read the international monetary expression or the local monetary expression.
- The following functions are new for SAS 9.2:
  - The GETPXLOCALE function returns the POSIX locale value for a SAS locale.
  - The GETPXREGION function returns the current, two-letter region code.
  - The SORTKEY function creates a linguistic sort key.
  - The UNICODELEN function creates a linguistic sort key.
  - The UNICODEWIDTH function specifies the length of a display unit for the Unicode data.
- The following system options are new for SAS 9.2 NLS:
  - The BOMFILE system option specifies whether to write the Byte Order Mark (BOM) prefix on Unicode encoded external files.
  - The RSASIOTRANSERROR system option displays a transcoding error when illegal data is read from a remote application.

For more information, see the complete list of changes in the SAS National Language Support (NLS): Reference Guide. Let us concentrate now on two major new features: Language Switching and Linguistic Sorting.

### **CHANGING LANGUAGE DURING A SAS SESSION**

By default, a SAS Unicode Server session in 9.2 has set the option LOCALELANGCHG. If you want to switch this option off, you need to add the following line to the SAS configuration file, which is located in !SASROOT/nls/u8: - NOLOCALELANGCHG.

Being able to switch the language of reports on the fly is ideal for countries that have more than one official language. Belgium, for example, has Dutch, French and German; Finland has Finnish and Swedish; and Switzerland has four national languages: German, French, Italian, and Rhaeto-Romanic. In Japan, English is the foremost foreign language studied and taught, so publications might appear in Japanese and in English. In all these cases, you can use various client applications in conjunction with a SAS Unicode Server session to produce multilingual reports on the fly. Below are some sample scenarios to illustrate the process.

## SCENARIO1: CONNECT WITH SAS® DATA INTEGRATION STUDIO TO THE SAS UNICODE SERVER

A simple test can be run by submitting some code to the server to run a Natural Language (NL) format. The NL formats are locale-sensitive and should return values that reflect the locale that is being set. See the SAS 9.2 National Language Support (NLS): Reference Guide for details on using the NL formats. In our example we also include a PROC OPTIONS step to verify the option values:

```
proc options group=languagecontrol; run;
data _null_;
d=today();
put d nldate.;
run;
```

In this case, we have set the locale for SAS<sup>®</sup> Data Integration Studio to French by using the Locale Setup Manager. The Locale Setup Manager is a Java client that enables users to configure the locale of SAS Java clients:

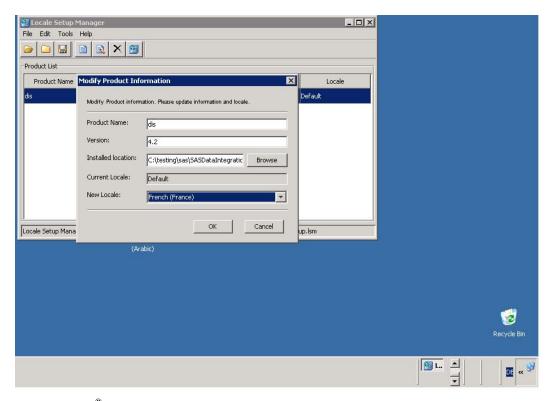


Figure 4: SAS® Locale Setup Manager 2.1

The result of our simple test program looks like this:

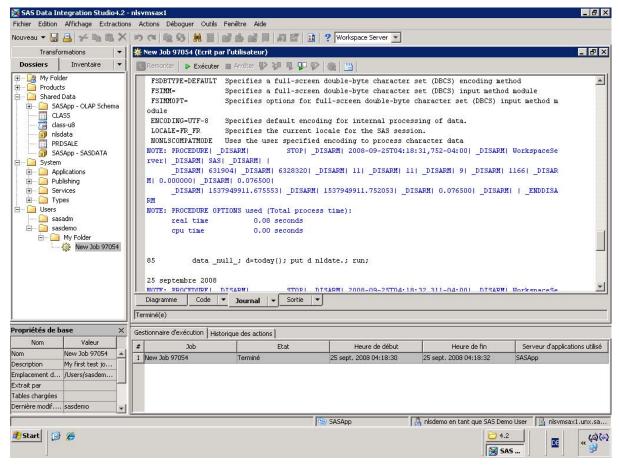


Figure 5: SAS® Data Integration Studio in French

We can see that the SAS Unicode Server session has picked up the language of the client. The date is also written out in the current locale.

### **SCENARIO 2: CREATE A BILINGUAL REPORT**

Here we are using SAS® Enterprise Guide® and the SAS Unicode Server to create a report in both English and Japanese. We are switching the language on the fly:

```
ods html file='bilingual-report.html';

options locale=en_US;
title "executed in &sysvlong by &sysuserid on machine &SYSTCPIPHOSTNAME";
footnote "%sysfunc(datetime(),nldatm.)";
proc univariate data=sashelp.class;
var age;
run;

options locale=ja_JP;
title "マシン &SYSTCPIPHOSTNAME.で、 &sysuserid によって &sysvlong で実行されました。";
footnote %sysfunc(datetime(),nldatm.);
proc univariate data=sashelp.class;
var age;
run;
ods html close;
```

This is what the result looks like:

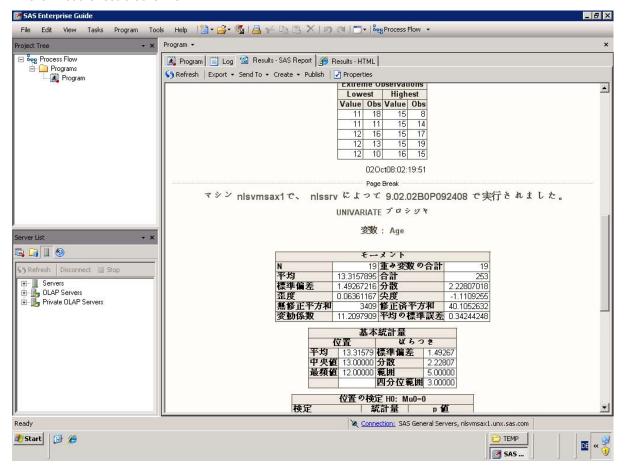


Figure 6: Bilingual Report (English - Japanese)

### SCENARIO 3: CREATE SEPARATE REPORTS DEPENDING ON THE LANGUAGE, AND USE LINGUISTIC SORTING

In this example we are assuming that we create lists of Swiss customers from a multilingual database:

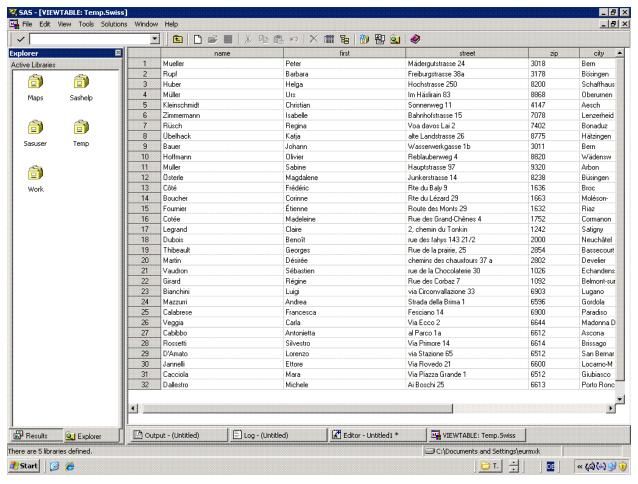


Figure 7: Multilingual Address Data

To accommodate users' language preferences, we create separate lists for French-, German-, and Italian-speaking customers. For output, we sort these lists according to local language conventions, using SORTSEQ=LINGUISTIC. Using this option produces a collating sequence that is culturally correct.

But what does this mean? Although there are recognized standards for collation, the way people look at data in "sorted" order differs a lot. German collation is different from French, and a Danish one is again different from both—just to name a few. To implement linguistic collation, SAS has adopted the International Components for Unicode (ICU). The ICU and its implementation of the Unicode Collation Algorithm (UCA) have become a de facto standard. In earlier versions of the SAS System, international customers were able to use translation tables (trantabs) for defining alternative collating sequences. Though this possibility still exists, it does have serious restrictions. Translation tables are limited to remapping or reordering up to 256 characters from a single-byte encoding. A trantab is limited in its ability to create a collating sequence that is intuitive or that meets cultural expectations. For example, French compares accented characters from right to left, not from left to right as in other languages. A sort using a translation table cannot make such a fine distinction, because a translation table can assign only a single "weight" to a character and cannot distinguish between several collation levels.

Table 1 shows an example of French names. In the first column, we see an unordered list of these names. In the second, we have used a translation table that orders the letters of the French alphabet, along with the accented variants, so that the lowercase and uppercase characters alternate and accented character variants are assigned weights close to the unaccented characters (for example, 'a' < 'A' < 'a' < 'Â' < 'â' < 'Â' < ...). However, the single-level

weighting approach used in trantab collation is not sufficient for properly ordering the list of words. The third column shows the same list of names sorted linguistically, according to the rules of the French language.

Unordered List	TRANTAB Collation	Linguistic Collation
	FRSOLAT1	LOCALE=fr_FR
Côté	Boucher	Boucher
Boucher	Cotée	Côté
Fournier	Côté	Cotée
Cotée	Dubois	Dubois
Legrand	Fournier	Fournier
Dubois	Girard	Girard
Thibeault	Legrand	Legrand
Martin	Martin	Martin
Vaudron	Thibeault	Thibeault
Girard	Vaudron	Vaudron

**Table 1: Linguistic Collation (French)** 

Now let us come back to creating separate lists for French-, German-, and Italian-speaking customers. Here's the code:

```
Create subsets of the data
  ----- * /
proc sql;
create table french as
select name, first, street, zip, city from temp.swiss where lang = 'fra';
create table german as
select name, first, street, zip, city from temp.swiss where lang = 'deu';
create table italian as
select name, first, street, zip, city from temp.swiss where lang = 'ita';
quit;
  label variables in different languages
  */
data french;
  set french;
  label name='Nom';
  label first='Prénom';
  label street='Adresse';
  label city='Lieu';
  label zip='CP';
run:
data german;
  set german;
  label name='Name';
  label first='Vorname';
  label street='Strasse';
  label city='Ort';
  label zip='PLZ';
run;
data italian;
  set italian;
  label name='Nome';
  label first='Cognome';
  label street='Strada';
  label city='Città';
  label zip='CAP';
```

```
run;
/* -----
  Run the SORT Procedure and create RTF output
ods rtf file='c:\temp\french.rtf';
options locale=fr_CH dflang=locale;
TITLE;
TITLE1 "Liste des clients";
TITLE2 "par ordre alphabétique";
FOOTNOTE:
FOOTNOTE1 "Générée par le système SAS le %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) à
%SYSFUNC(TIME(), NLTIMAP10.)";
proc sort data=french sortseg=linguistic;
 by name;
run;
proc print data=french label noobs; run;
RUN; QUIT;
TITLE: FOOTNOTE:
  DTRESET specifies that SAS update the date and time in the titles
  of the SAS log and the listing file.
  -----*
OPTIONS DTRESET;
ods rtf file='c:\temp\german.rtf';
options locale=de_CH dflang=locale;
TITLE1 "Kundenliste";
TITLE2 "alphabetisch sortiert";
FOOTNOTE;
FOOTNOTE1 "Generiert durch das SAS System am %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) um
%SYSFUNC(TIME(), NLTIMAP10.)";
proc sort data=german sortseq=linguistic;
 by name;
run:
proc print data=german label noobs; run;
RUN; QUIT;
TITLE; FOOTNOTE;
              _____
  DTRESET specifies that SAS update the date and time in the titles
  of the SAS log and the listing file.
   -----*/
OPTIONS DTRESET;
ods rtf file='c:\temp\italian.rtf';
options locale=it_CH dflang=locale;
TITLE1 "Lista dei clienti";
TITLE2 "in ordine alfabetico";
FOOTNOTE;
FOOTNOTE1 "Generato dal sistema SAS il %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) alle
%SYSFUNC(TIME(), NLTIMAP10.)";
proc sort data=italian sortseq=linguistic;
 by name;
run:
proc print data=italian label noobs; run;
RUN; QUIT;
TITLE; FOOTNOTE;
ods rtf close;
```

ኛ SAS - [Results Viewer - italian.rtf] \_ B × File(F) Edit(E) View(V) Go(G) Tools(T) Solutions(S) Window Help(H) \_ B × ~ ▼ | Results 🚊 👍 Print: Liste des clients Table WORK.FRENCH
Table WORK.FRENCH lundi, 13. octobre 2008 07. 🖃 🔓 Print: Kundenliste 🔛 Datei WORK.GERMAN 🖷 Datei WORK.GERMAN Liste des clients Print: Lista dei clienti

Data set WORK.ITALIA

Data set WORK.ITALIA par ordre alphabétique Nom Prénom CP Lieu Adresse 1663 Moléson-sur-Gruyères Boucher Corinne Rte du Lézard 29 Côté 1636 Broc Frédéric Rte du Baly 9 Cotée Madeleine Rue des Grand-Chênes 4 1752 Cormanon Dubois Benoît rue des fahys 143 21/2 2000 Neuchâtel 1632 Riaz Fournier Route des Monts 29 Girard Régine Rue des Corbaz 7 1092 Belmont-sur-Lausanne Legrand Claire 2, chemin du Tonkin 1242 Satigny Martin Désirée chemins des chauxfours 37 a | 2802 | Develier Thibeault Georges Rue de la prairie, 25 2854 Bassecourt Vaudron Sébastien rue de la Chocolaterie 30 1026 Echandens ■ 3 ■ 3 0 4 Log - (Untitled) Editor - Untitled1 \* 🗗 Results **Explorer** Results Viewer - italia... C:\Documents and Settings\eurmxk 🏄 Start 🏻 🍪 🍎 **3** U. (4)(4) DE 雾 5..

And this is what the list of French-speaking customers looks like:

Figure 8: Report in RTF Format

### **ACCESSING AND PROCESSING DATA FROM VARIOUS SOURCES**

A SAS Unicode Server session is ideal for processing data from various sources. Data can be read into the session from three sources:

- External files
- SAS libraries
- DBMS tables

SAS 9.2 has enhanced the accessing of external files and DBMS tables. As an example, let us look at the SAS/ACCESS $^{\otimes}$  Interface to PC Files product.

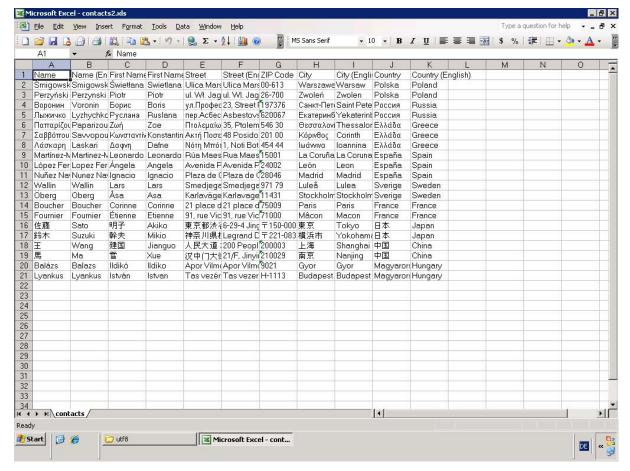


Figure 9: Spreadsheet with Multilingual Address Data

In SAS 9.2, a LIBNAME option, UNICODE=YES | NO, is available for the SAS/ACCESS, Excel, and PCFILES LIBNAME engines.

If you specify UNICODE=YES, SAS binds all text data in wide-character format and transcodes character data to Unicode data that can be in a Microsoft Access file or in an Excel file. If you are running SAS in a Windows UTF-8 session, the SAS/ACCESS and Excel LIBNAME engines automatically set UNICODE=YES. However, you have to specify this LIBNAME option when you use the PCFILES LIBNAME engine.

The following code imports multilingual data from a spreadsheet into a UTF-8 data set:

```
proc import out=work.contacts
  datafile= "c:\utf8\contacts.xls"
  dbms=excel replace;
  range="contacts";
  getnames=yes;
  mixed=no;
  scantext=yes;
  usedate=yes;
  scantime=yes;
run;
```

This works only if you have file-system-level access to the Excel file.

If you need to access the same spreadsheet from a UNIX machine, you need a LIBNAME statement that specifies the PCFILES engine. An example would be:

```
libname myxls pcfiles server="ipaddress" type=excel port=8621
    path="c:\utf8\contacts.xls" unicode=yes;
```

You do not need to specify "type=excel". If the file extension is xls (or xlsb, xlsx, xlsm for Excel 2007), type=excel is automatically assumed. The same goes for .mdb and .accdb extensions. They assume type=access. Also, if the default port=8621 is used, you can omit the port option. However, if the PC Files server is started with a different, non-standard port number, then the PORT option must be specified.

Please verify or keep in mind the following details:

- The PC Files Server is installed on the 32-bit Windows system, and the files that you want to read from are also located here, or there is file-system-level network access with something like PATH=\\server\my\_dir\mysheet.xls. The PC Files Server can also run on a 64-bit Windows system, but it will run in 32-bit compatibility mode.
- The SAS/ACCESS Interface to PC Files is installed on the UNIX or LINUX system.
- The PC Files Server must be started before you can access files located in Windows PC from UNIX or LINUX.
- UNIX or LINUX users must know the following details before they can access Windows files:
  - the server name, which means the Windows short name.
  - the service port used by the PC Files Server. You can change the port as you wish in the PC Files Server; after the change, restart the PC Files Server to make the change take effect.
  - the exact file path that you want to access.

### CONCLUSION

This paper, combined with the recommended reading, should provide you with basic information about the new language features and enhancements for SAS 9.2. We suggest that you explore the new features and enhancements. You can gain enhanced understanding by using these new features as you develop or use a SAS solution.

### **REFERENCES**

SAS Institute Inc. 2008. SAS® 9.2 National Language Support (NLS): Reference Guide. Cary, NC: SAS Institute Inc.

### **ACKNOWLEDGMENTS**

I wish to thank the employees here at SAS for their contributions and constructive comments. A special thank you goes to Chris Hemedinger for permission to use material from his blog entry "Babelfishing in your SAS 9.2 programs."

#### RECOMMENDED READING

- SAS Institute Inc. 2008. SAS Institute technical paper. "Linguistic Collation: Everyone Can Get What they Expect -Sensible Sorting for Global Business Success."
- SAS Institute Inc. 2008. SAS Institute technical paper. "Processing Multilingual Data with the SAS® 9.2 Unicode Server."
- SAS Institute Inc. 2008. SAS Institute technical paper. "Changing language during a SAS session."

### **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at Manfred.Kiefer@sas.com.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.