L'ANALYSE DE LA FREQUENCE DES ENCHAINEMENTS

Munie des arguments présentés dans le tableau ci-dessous, la macro-expression « chaines » recherche, dans le tableau d'entrée *filin*, les chaînes de deux mots consécutifs puis calcule le nombre, la fréquence, le rang et l'indice Pareto (produit du rang par la fréquence) de chaque chaîne de caractères.

Argument	Signification
ligne	Nom symbolique Sas d'une variable chaîne de caractères.
filin	Nom symbolique du tableau d'entrée (Data) Sas qui contient la chaîne de caractères *ligne*.
filout	Nom symbolique du tableau (Data) Sas de sortie.

Dès le début de son exécution, la macro-expression « chaines » transforme au moyen de la fonction TRANSLATE ci-dessous toutes les lettres accentuées en lettres non accentuées majuscules. La forme conjuguée « a » de l'auxiliaire « avoir » est ainsi assimilée à la préposition « à ». Il est évident que vous pouvez en décider autrement en modifiant l'ordre TRANSLATE.

```
%MACRO chaines (ligne,filin,filout);
DATA travail;
SET &filin;
%traduis(&ligne);
&ligne=TRANSLATE(&ligne ,'AAAEEEEIIOOUUU','àäâéèëêïîöôùüû');
&ligne=UPCASE(&ligne);
i=1;
debut : mot = SCAN (&ligne,i,"$()&:'-!.,;/? ");
mot = COMPRESS(mot);
mot=TRANSLATE (mot,'000000000', '123456789');
IF INDEX(mot,'0') NE 0 THEN GOTO finligne;
IF mot=' ' THEN GOTO finligne;
ELSE DO;
  molag=LAG1(mot);
  chaine = compress(molag !! '_' !! mot) ;
  OUTPUT;
  i=i+1;
  GOTO debut ;
  END;
finligne :
PROC FREQ NOPRINT ;
TABLES chaine /OUT=travail ;

/* Ajout du rang et de l'indice pareto ; */

PROC SORT DATA=travail ;
BY DESCENDING COUNT;
DATA &filout ;
SET travail ;
RETAIN numrang 0;
IF COUNT NE LAG1(COUNT) THEN numrang=numrang+1;
pareto=COUNT*numrang;
%MEND ;
```

Adib Gabriel Hathout – Copyright Mars 2002

Exemple d'utilisation de la macro-expression « chaines »

Dans le tableau d'entrée Sas *madata* qui contient un flux discursif sur l'actualité internationale de janvier 2001 à octobre 2001, il existe une variable chaîne de caractères « textes » à laquelle on peut appliquer la macro-commande %chaines comme suit :

```
%chaines (textes, madata, resultat) ;
RUN ;
```

A l'image du tableau (Data) de sortie *resultat ci-dessous*, les chaînes de caractères rencontrées sont présentées par nombre décroissant (COUNT) avec la fréquence POURCENT, le rang NUMRANG et l'indice PARETO de chaque chaîne de caractères. Toutes les chaînes n'ont pas la même signification linguistique. Ainsi, les chaînes comme DE LA et A LA relèvent de la nature de la langue française alors qu'une chaîne comme « Jacques Chirac » relève du langage. Une éventuelle transformation du langage ne peut être détectée après une seule exécution de la macro-commande « chaines » : deux exécutions sont au minimum nécessaires pour détecter l'éventuelle évolution du langage.

The SAS System 1
11:45 Friday, February 1, 2002

OBS	CHAINE	COUNT	PERCENT	NUMRANG	PARETO
1	DE_LA	356	0.96642	1	356
2	DE_L	336	0.91213	2	672
3	A_LA	168	0.45606	3	504
4	AUJOURD_HUI	137	0.37191	4	548
5	ET_PUIS	127	0.34476	5	635
6	A_L	102	0.27690	6	612
7	BONJOUR_A	96	0.26061	7	672
8	L_ACTUALITE	89	0.24160	8	712
9	C_EST	88	0.23889	9	792
10	A_TOUS	83	0.22532	10	830
11	ETATS_UNIS	80	0.21717	11	880
12	EN_FRANCE	78	0.21174	12	936
13	CE_MATIN	77	0.20903	13	1001
14	DANS_LE	74	0.20088	14	1036
15	IL_Y	72	0.19546	15	1080
16	Y_A	72	0.19546	15	1080
17	LE_PRESIDENT	69	0.18731	16	1104
18	A_ETE	68	0.18460	17	1156
19	DANS_L	64	0.17374	18	1152
20	DANS_LA	63	0.17102	19	1197
21	D_UN	60	0.16288	20	1200
22	SUR_LE	58	0.15745	21	1218
23	DANS_UN	56	0.15202	22	1232
24	DANS_LES	55	0.14931	23	1265
25	PROCHE_ORIENT	54	0.14659	24	1296
26	LA_UNE	52	0.14116	25	1300
27	SUR_LA	49	0.13302	26	1274
28	UNE_DE	48	0.13030	27	1296
29	AU_PROCHE	47	0.12759	28	1316
30	HIER_SOIR	45	0.12216	29	1305
31	JACQUES_CHIRAC	44	0.11945	30	1320
32	SUR_L	44	0.11945	30	1320
33	S_EST	44	0.11945	30	1320
34	CE_JOURNAL	43	0.11673	31	1333
35	POUR_LA	42	0.11402	32	1344
36	PREMIER_MINISTRE	42	0.11402	32	1344
37	SUR_LES	41	0.11130	33	1353
38	D_UNE	40	0.10859	34	1360
39	LA_FRANCE	40	0.10859	34	1360
40	DANS_UNE	38	0.10316	35	1330

L'ordre trouvé peut ensuite être conforté ou infirmé en étudiant la liaison entre le nombre et le rang de chaque chaîne de caractères avec la procédure Proc REG; Model Count=Numrang ;

Adib Gabriel Hathout – Copyright Mars 2002