L'ANALYSE DE LA RESONANCE

D'UNE STRUCTURE LEXICALE

Les macro-expressions %retentis et %retent comparent les structures lexicales de deux textes T1 et T2 (lemmatisés ou non) supposés ici être des tableaux (Data) Sas. On peut aussi bien s'intéresser à la résonance de T1 par rapport à T2 ou à la résonance de T2 par rapport à T1. Le résultat est un ensemble de mots qui caractérisent, selon le cas, T1 par rapport à T2 ou T2 par rapport à T1.

Les macro-expressions %retentis et %retent détectent les mots qu'il faut souligner dans un texte T1 lorsqu'il est comparé à un texte T2. Les macro-expressions %retent et %retentis impriment, au plus 45 mots, typiques du texte où l'on veut souligner des mots. Le nombre de 45 mots est ramené à un nombre nb inférieur à 45 lorsque la résonance n'est pas assez significative, le seuil de significativité étant calculé par le programme lui-même au regard de l'écart-type des proportions de fréquence entre les deux textes.

La macro-variable *delimite*, contenue dans chaque macro-expression %retent et %retentis, est supposée contenir les délimiteurs. Le plus simple est ici de la définir dans le texte de votre programme (par exemple en écrivant: %LET delimite = '.' ;). Sur P.C., vous pouvez aussi définir les délimiteurs dans votre AUTOEXEC.SAS, et donner - le cas échéant - un contrordre pour une application particulière.

La macro – expression %retentis

Pour comparer les structures lexicales de deux textes, il faut d'abord calculer, pour chaque texte, sa propre structure lexicale. C'est ce que fait la macro-expression %freqsas dont les arguments figurent dans le tableau ci-dessous.

Argument	Signification
filin	Nom symbolique du tableau (Data) d'entrée Sas qui contient le texte
ligne	Nom d'une variable chaîne de caractères contenue dans *filin*
filout	Nom symbolique du tableau (Data) Sas créé par %freqsas pour contenir les structures lexicales du texte.

Sur la base du tableau (Data) *filin* lu en entrée, la macro–expression *freqsas* ci-dessous enregistre dans le tableau de sortie *filout* les deux variables suivantes : la variable *mot* (un mot ne contenant pas de chiffres revenant au moins une fois dans le texte) et la variable *Count* (le nombre de fois que le mot revient),les différents mots sélectionnés étant finalement triés par ordre décroissant sur la base de la variable Count.

```
%MACRO freqsas (ligne,filin,filout);

DATA &filout;
SET &filin ;
DROP &ligne;
i=1;
debut : mot = SCAN (&ligne,i,&delimite);
IF mot=' ' THEN GOTO finligne;
ELSE DO;
  mot=COMPRESS(mot) ;
  mot=UPCASE(mot);
  OUTPUT ;
  i=i+1;
  GOTO debut ;
  END;
finligne :

PROC FREQ NOPRINT ;
TABLES mot/OUT=&filout;

PROC SORT;
BY COUNT;

%MEND ;
```

Munie des quatre arguments présentés dans le tableau ci-dessous, la macro-expression %retentis fait appel deux fois au programme *freqsas* qui analyse successivement le tableau d'entrée Sas (Data) *ficimage,* puis le tableau d'entrée Sas (Data) de référence *ficappar*.

Argument	**Signification**
ficimage	Nom du tableau (Data) Sas qui contient le texte dont on veut analyser les résonances. Le texte est supposé contenu dans la variable chaîne de caractères "textes".
ficappar	Nom du tableau (Data) Sas qui contient le texte par rapport auquel les résonances du tableau Sas *ficimage* émergent. Le texte est supposé contenu dans la variable chaîne de caractères "textes".
sortie	Nom du tableau Sas (Data) contenant les résultats de l'analyse.
limite	Seuls les mots qui surviennent un nombre de fois au moins égal à "limite" sont pris ici en considération.

A la fin de l'exécution de la macro-commande %rententis figurant à la page suivante, le tableau Sas (Data) *sortie* présente quatre variables : la variable *mot* contenant des mots qui reviennent un nombre de fois au moins égal à l'argument "limite", la variable *nbimag* recensant le nombre de fois qu'un mot revient dans le fichier "image" *ficimage*, la variable

nbappar rappelant le nombre de fois que le même mot revient dans le tableau Sas de référence *ficappar* et la variable *forimage*, rapport entre les nombres *nbimag* et *nbappar*.

```
%MACRO retentis (ficimage , ficappar , sortie, limite ) ;
%freqsas (textes, &ficimage, travimag);
%freqsas (textes, &ficappar, travappa);
PROC SORT DATA=travimag ; BY mot;
PROC SORT DATA=travappa ; BY mot;
DATA travimag;
SET travimag;
nbimag=COUNT;
KEEP nbimag mot;
DATA travappa;
SET travappa;
nbappa=COUNT;
KEEP nbappa mot;
DATA &sortie;
MERGE travimag travappa; BY mot;

/* Calcul de forimage et élimination des mots contenant un chiffre */
DATA &sortie;
SET &sortie;
forimage=nbimag/nbappa;
mot=TRANSLATE (mot,'000000000','123456789');
IF INDEX (mot,'0') NE 0 THEN DELETE;
IF forimage NE .;
%traduis(mot);
PROC SORT DATA=&sortie;
BY DESCENDING forimage;

/* Calcul de l'écart-type de la variable forimage (Force de l'image) */
PROC SUMMARY NWAY DATA=&sortie;
VAR forimage;
OUTPUT OUT=travail STD=s;
DATA _NULL_;
SET travail END=toto;
IF toto=1 THEN CALL SYMPUT ('ecart', s);

/* Conservation du fichier intégral des caractérisations */
DATA travail ;
SET &sortie ;
seuil = &ecart;

/* Définition de la DATA SAS &sortie */
DATA &sortie ;
SET travail;
IF forimage >=seuil;
IF (nbimag+nbappa) >= &limite;
IF _N_ <= 45 ;

/* Résultats finaux abrégés sur une seule page */
TITLE1 "45 premiers signaux retentissants dans la table SAS &ficimage" ;
TITLE2 " relativement à la table SAS &ficappar " ;
TITLE3 "revenant plus de &limite fois avec Seuil de significativité >=
&ecart";
TITLE4 "Résultats intégraux dans work.travail";

PROC CHART DATA=&sortie;
HBAR mot/SUMVAR=forimage DESCENDING;

%MEND;
```

La macro – expression %retent

Munie des mêmes arguments que la macro-expression retentis, la macro-commande %retent fait également appel deux fois au programme *freqsas* qui analyse successivement le tableau d'entrée Sas (Data) *ficimage,* puis le tableau d'entrée Sas (Data) de référence *ficappar*.
Les deux macro-expressions %retentis et %retent se différencie uniquement par le traitement réservé aux mots non apparus dans l'un des tableaux de sortie de la macro-expression *freqsas*. Dans la macro-expression *retent*, les données manquantes des variables *nbimag* et *nbappa* sont ainsi supposées égales à zéro ce qui entraîne l'ajout d'une unité au dénominateur de la variable *forimage* (au cas où surviendrait une valeur nulle pour *nbappa*).

```
%MACRO retent (ficimage , ficappar , sortie, limite ) ;
%freqsas (textes, &ficimage, travimag);
%freqsas (textes, &ficappar, travappa);
PROC SORT DATA=travimag ; BY mot; PROC SORT DATA=travappa ; BY mot;
DATA travimag; SET travimag; nbimag=COUNT; KEEP nbimag mot;
DATA travappa; SET travappa; nbappa=COUNT; KEEP nbappa mot;
DATA &sortie; MERGE travimag travappa; BY mot;
DATA &sortie; SET &sortie ;
IF nbappa=. THEN nbappa=0; IF nbimag=. THEN nbimag=0;

/* Calcul de forimage et élimination des mots contenant du chiffre
   Le +1 dans le dénominateur est destiné à éviter la division par
zéro*/

DATA &sortie; SET &sortie; forimage=nbimag/(nbappa+1);

mot=TRANSLATE (mot,'000000000','123456789');
IF INDEX (mot,'0') NE 0 THEN DELETE;IF forimage NE .;%traduis(mot);
PROC SORT DATA=&sortie; BY DESCENDING forimage;

/* Calcul de l'écart-type de la variable forimage (Force de l'image)
*/

PROC SUMMARY NWAY DATA=&sortie; VAR forimage; OUTPUT OUT=travail
STD=s;
DATA _NULL_; SET travail END=toto;
IF toto=1 THEN CALL SYMPUT ('ecart', s);

/* Conservation du fichier intégral des caractérisations */
DATA travail ; SET &sortie ; seuil = &ecart;

/* Extraction de la table SAS requise définie par l'argument &sortie
*/

DATA &sortie ; SET travail;
IF forimage >=seuil; IF (nbimag+nbappa) >= &limite; IF _N_ <= 45 ;

/* Résultats finaux abrégés sur une seule page */

TITLE1 "45 premiers signaux retentissants dans la table SAS
&ficimage" ;
TITLE2 " relativement à la table SAS &ficappar " ;
TITLE3 "revenant plus de &limite fois avec Seuil de significativité
>= &ecart";
TITLE4 "Résultats intégraux dans work.travail";
PROC CHART DATA=&sortie; HBAR mot/SUMVAR=forimage DESCENDING;
%MEND;
```

Exemple de résultats

Le service d'étude du discours politique international de la bibliothèque de documentation internationale contemporaine a tenté de souligner, à l'aide de la macro-expression %retentis, les mots les plus fréquents du discours politique international tenu au cours de la période comprise entre le 11 septembre 2001 et le 25 octobre 2001. L'outil %retentis a été utilisé ici de deux différentes manières :

1° - en considérant que le tableau Sas de référence ("ficappar ") est le discours sur l'actualité tenu entre le 12 septembre 2000 et le 11 septembre 2001.

2° - en considérant que le tableau Sas de référence (ficappar) est le discours sur l'actualité tenu entre le 12 septembre 1997 et le 11 septembre 2001.

Ainsi le tableau de référence qui permet de souligner les mots importants du discours sur l'actualité est visualisé par rapport à l'année écoulée dans le premier cas, et par rapport aux trois années écoulées dans le second cas.

Les programmes correspondant présents respectivement dans les cadres n°1 et n°2 ci-après, sont suivis par les résultats qu'a donné la macro-expression *%retentis*. La même période d'analyse permet de constater une évolution de la variable *forimage* bien moins brutale lorsque nous prenons du recul.

Premières valeurs prise par la variable *forimage* dans l'histogramme N°1 : 31, 17, 6, 6, 6, 6, …

Premières valeurs prises par la variable *forimage* dans l'histogramme N°2 : 5.6, 3, 3, 3, 2.8, …

Une fois les mots soulignés par l'expression *%retentis*, la question focale que pose l'utilisation de ces modules dans le contexte de l'informatique décisionnelle est, bel et bien, celle de la décision; en l'occurrence la décision de lecture. Le retour au contexte phraséologique qui distingue les mots soulignés dans l'histogramme numéro 2 et non soulignés dans le précédent répond à la question « Qu'est-ce qui émerge lorsque nous prenons du recul ? ». La réponse est dans des termes comme *tourisme* qui nous renvoie à l'assassinat du ministre israélien du tourisme.

Plus globalement, les avantages de l'utilisation interactive peuvent difficilement être illustrés sur le papier. Dans l'idéal,l'utilisateur doit intégrer la macro-expression *retentis* dans sa propre problématique et dans ses propres bases de données.

Cadre N°1

```
DATA new old;
set basactu.noyau;
IF 20010911 <= date <= 20011025 then output new;
IF 20000912 <= date <= 20010911 then output old;
%retentis (new, old, sortie, 1);
RUN;
```

HISTOGRAMME 1

MOT		Impact
AFGHANISTAN	,******************************	31.00000
KABOUL	,****************	17.00000
TALIBANS	,*******	6.50000
FRAPPES	,******	6.00000
BOMBARDEMENTS	,******	6.00000
AFGHANE	,******	6.00000
PHASE	,*****	5.00000
PAKISTAN	,*****	5.00000
MASSOUD	,*****	5.00000
KRACH	,*****	5.00000
PENTAGONE	,****	4.00000
LIVRER	,****	4.00000
DEUIL	,****	4.00000
UTILISE	,***	3.00000
SOL	,***	3.00000
MYSTERE	,***	3.00000
MILLIARDAIRE	,***	3.00000
ISLAMABAD	,***	3.00000
GALATASARAY	,***	3.00000
ATTENTATS	,***	2.87500
REUNIT	,**	2.00000
REPUBLICAINE	,**	2.00000
OMAR	,**	2.00000
NOBEL	,**	2.00000
MEURTRIERS	,**	2.00000
MASSACRE	,**	2.00000
LANGUE	,**	2.00000
ISLAM	,**	2.00000
INCLINE	,**	2.00000
GRIEVEMENT	,**	2.00000
GRAVEMENT	,**	2.00000
FRANCHI	,**	2.00000
EXPOSE	,**	2.00000
ESSAYE	,**	2.00000
ENTRETENU	,**	2.00000
DETRUIT	,**	2.00000
CONTINUENT	,**	2.00000
COMMENTER	,**	2.00000
CLAIR	,**	2.00000
CAMIONS	,**	2.00000
AVIONS	,**	2.00000
AVIATION	,**	2.00000
APPAREILS	,**	2.00000
AMERICAINES	,**	2.00000

Šƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒ
5 10 15 20 25 30

Cadre N°2

```
DATA new old;
SET basactu.noyau;
IF 20010911 <= date <= 20011025 THEN OUTPUT new;
IF 19970912 <= date <= 20010911 THEN OUTPUT old;
%retentis (new, old, sortie, 1);
RUN;
```

```
                          HISTORGRAMME 2
     MOT                                                  Impact

     KABOUL          ,****************************       5.666667
--> TOURISME         ,***************                   3.000000
     ISLAMABAD       ,***************                   3.000000
     AFGHANE         ,***************                   3.000000
--> OUSSAMA          ,**************                    2.833333
--> BENLADEN         ,**************                    2.750000
     MASSOUD         ,************                      2.500000
     AFGHANISTAN     ,***********                       2.214286
--> USA              ,**********                        2.000000
--> TUPOLEV          ,**********                        2.000000
--> TODAY            ,**********                        2.000000
--> STRUCTURES       ,**********                        2.000000
--> PSYCHOSE         ,**********                        2.000000
--> MOLLAH           ,**********                        2.000000
--> MIRE             ,**********                        2.000000
     GELE            ,**********                        2.000000
--> FBI              ,**********                        2.000000
     EXPOSE          ,**********                        2.000000
--> ELIMINES         ,**********                        2.000000
--> DETOURNES        ,**********                        2.000000
--> COURRIER         ,**********                        2.000000
--> ALERTES          ,**********                        2.000000
                    Šƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒƒƒƒˆƒƒƒ
                         1    2    3    4    5
```

Note
====

--> pointe sur un mot qui émerge dans l'histogramme numéro 2 sans émerger dans l'histogramme numéro 1.

Précautions à prendre dans l'utilisation de %retentis et de %retent

Dans des textes qui ne sont pas assez longs, il est possible d'obtenir des résultats non significatifs. Dans ce cas, il est souvent utile, en cas de doute, de jeter un œil sur la manière dont les articles arrivent à une proportion stable au fur et à mesure que nous parcourons le texte. Ce travail peut être effectué par l'expression *%articles* ci-dessous muni de l'argument *filin* représentant le nom symbolique du tableau (Data) Sas qui contient le texte analysé dans une variable chaîne de caractères dénommée « textes ». A la sortie de la macro-expression %articles, le tableau (Data) de sortie Sas *travail* cumule les fréquences des articles (le, la, les, un, une, des) au fur et à mesure de la lecture de la variable chaîne « textes ».

Un texte ne peut pas cependant être tenu pour recevable du seul fait de la convergence des articles vers une proportion stable; le sujet ne pouvant pas être étudié indépendamment du contexte et du style. C'est pourquoi, il est aussi parfois utile d'observer la manière dont les pronoms relatifs (qui, que quoi, dont, où) convergent vers des proportions stables. La macro-commande %articles, donnée ci-dessous, doit vous aider dans l'interprétation en cas d'indécision.

```
%MACRO articles (filin);
DATA travail;
SET &filin ;
%MACRO mesvar ;
les las less uns unes dess ;
%MEND;
KEEP mot %mesvar;
RETAIN %mesvar 0 ;
i=1 ;
debut : mot=SCAN(textes,i,&delimite);
IF mot=' ' THEN GOTO finligne;
ELSE DO ;
mot=UPCASE(COMPRESS(mot));
IF mot='LE'   THEN DO; les+1;  OUTPUT; END;
IF mot='LA'   THEN DO; las+1;  OUTPUT; END;
IF mot='LES'  THEN DO; less+1; OUTPUT; END;
IF mot='UN'   THEN DO; uns+1;  OUTPUT; END;
IF mot= 'UNE' THEN DO; unes+1; OUTPUT; END;
IF mot= 'DES' THEN DO; dess+1; OUTPUT; END;
i=i+1 ;
GOTO debut;
END;
finligne:
DATA travail ;
SET travail;
KEEP ordre %mesvar;
ordre = _N_;
/* Profil des lignes */
ARRAY arvar(*) %mesvar ;
somvar=SUM(OF %mesvar);
DO i=1 TO DIM(arvar);
arvar(i)=arvar(i)/somvar;
END;
PROC PLOT  ;
PLOT ordre*las='*' ordre*uns='.'/OVERLAY;
TITLE 'Convergence des proportions de la (*) et (un(.) ';

%MEND;
```