mai 2017

Outil de statistique textuelle

R.TeMiS
Plugin de R Commander

Bénédicte Garnier, Institut national d'études démographiques (INED), F-75020 Paris, France

http://rtemis.hypotheses.org/

Ce document ne remplace pas un guide d'utilisation du logiciel mais montre une utilisation « pas à pas » et le paramétrage des menus avec un exemple de corpus.

Sommaire[1]

Tables des figures

[1] Je remercie Milan Bouchet-Valat et Elodie Baril pour leur relecture.

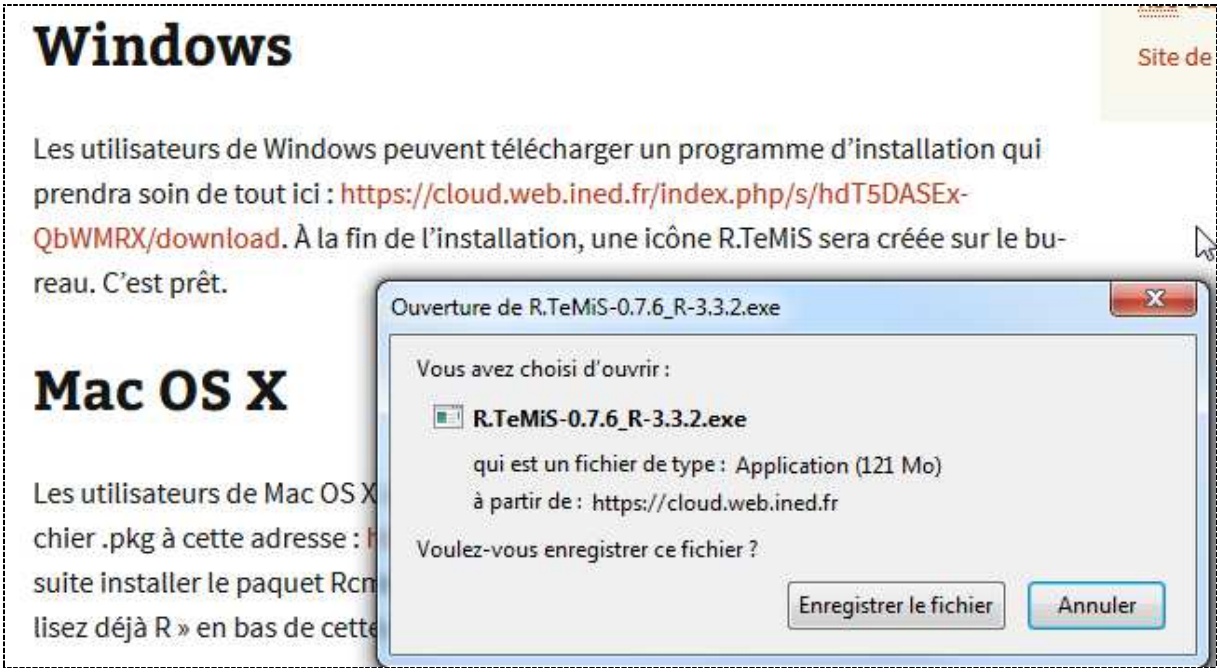
Installer R.TeMiS

Si R est déjà installé sur votre ordinateur, il faut le charger et lancer l'installation du plugin depuis le gestionnaire de packages ou avec les instructions suivantes (connexion à Internet requise) :

install.packages ("RcmdrPlugin.temis")

Si non, il faut le télécharger depuis http://rtemis.hypotheses.org/installation

Sous Windows, **le plus simple est** de télécharger le programme d'installation en cliquant sur le lien https://cloud.web.ined.fr/index.php/s/hdT5DASExQbWMRX/download. Il permet d'installer **en une seule fois**, R et tous les packages nécessaires à R.TeMiS



Il suffit ensuite de lancer l'exe et de suivre les instructions écran par écran.

Tutoriel de R.TeMiS

Se référer à la page fonctionnalités sur http://rtemis.hypotheses.org/r-temis-pas-a-pas.

Toutes les opérations sont documentées dans le bouton **Aide** des fenêtres de paramétrage des procédures.

Voir aussi : Milan Bouchet-Valat et Gilles Bastin, « RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R », The R Journal, 5 (1), 2013, p. 188-196.

La démarche d'analyse

Avant d'utiliser les menus de R.TeMiS, il est important de se situer dans la démarche d'analyse d'un corpus au moyen des méthodes de la statistique textuelle.

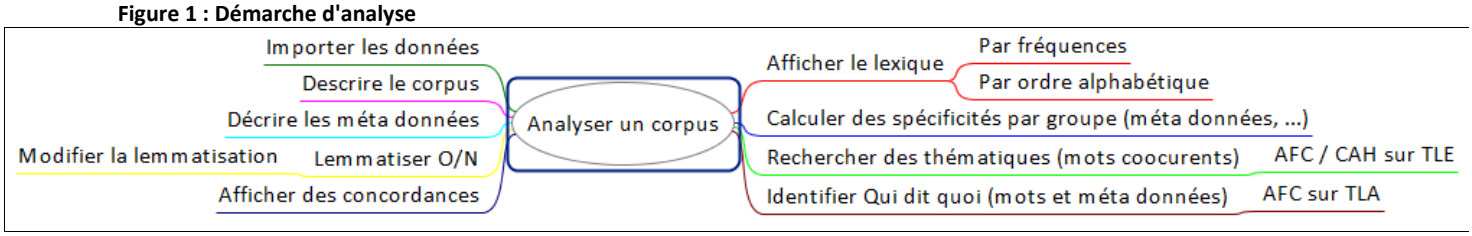
Après avoir importé ses données en spécifiant ses choix sur le type de corpus (texte ou tableur), la langue, la lemmatisation, le traitement des mots-outils ou des chiffres (importation du corpus) il faut « explorer » ses données avec des calculs de fréquences de « mots » (Afficher le lexique), des tris à plat et tris croisés sur les variables qualitatives (métadonnées).

Selon la problématique à traiter, il conviendra d'utiliser les méthodes adaptées pour rechercher les co-occurrences (analyses factorielles ou classifications), calculer des distances entre mots (dissimilarités), lister les mots spécifiques (spécificités), ou traiter ensemble les mots et les métadonnées (analyses factorielles).

L'analyse des concordances est également une aide à l'interprétation très précieuse.

R.TeMiS permet de créer des sous corpus ou de supprimer des mots (gestion du corpus) et de relancer de nouvelles analyses rapidement.

Figure 1 : Démarche d'analyse



Les données utilisées dans ce document sont extraites du projet EuroBroadMap (http://www.eurobroadmap.eu/). Nous traitons les réponses des étudiants interrogés en Chine à une question posée comme suit : « *Quels sont les mots que vous associez le plus à l'« Europe » ? Choisissez 5 mots au maximum ».*

Encadré 1 : Pourquoi choisir R.TeMiS

Ce plugin de RCommander permet à un débutant de s'initier à la programmation R et aux fonctions du package tm (text mining). Le script est généré au fur et à mesure du paramétrage des fenêtres correspondantes aux principales méthodes de la statistique textuelle. Les résultats et graphiques peuvent être exportés.

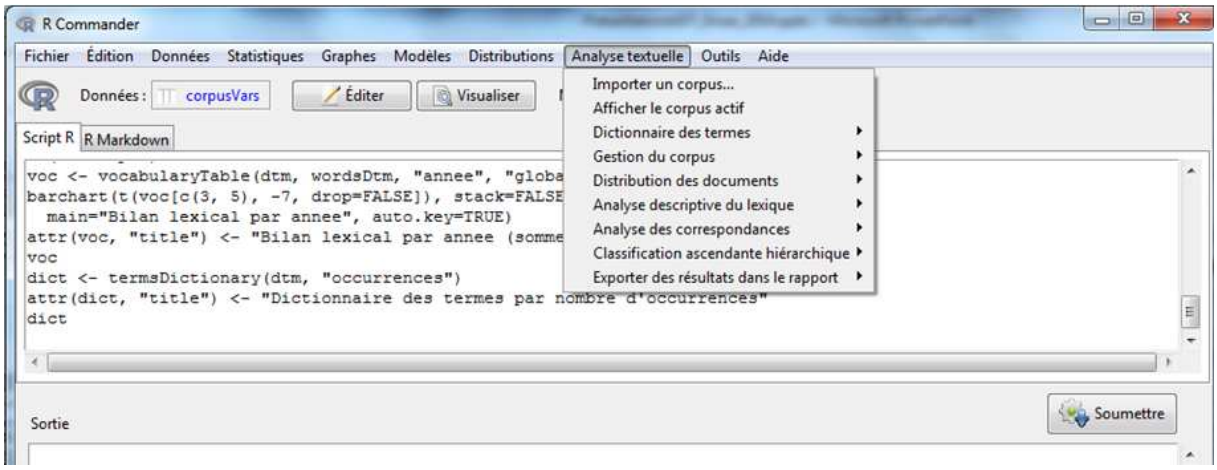
Le script R peut être intégré dans des routines et être sauvegardé pour réplication.

Lancer R.TeMiS

Exécuter la commande : **library (RcmdrPlugin.temis)** à partir de la **fenêtre RGui** (et pas R Studio).

Ou (suivant l'installation et votre environnement) directement à partir de l'**icône** :

Ici, nous utiliserons les différentes rubriques du menu **Analyse textuelle**[2] de R Commander



Importer un corpus constitué de textes « courts »

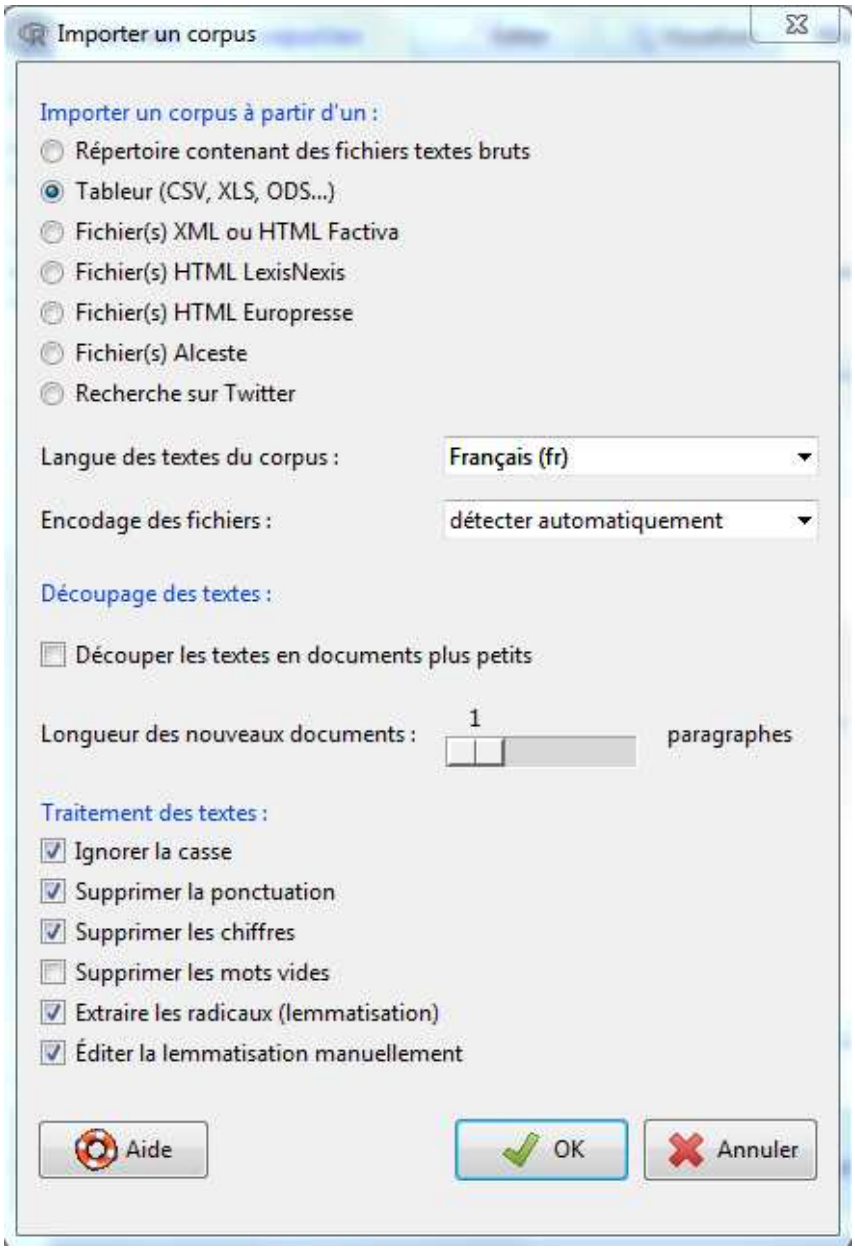
Dans notre exemple, les réponses sont constituées de quelques mots, on peut donc parler de textes courts[3]. C'est aussi le cas de réponses à des questions ouvertes ou de titres d'articles.

Figure 2 : Extrait du tableau contenant les textes courts à analyser et les variables qualitatives associées (métadonnées)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	D2_agr	G_City	G_Study	A1_Gender	AX_Nb	A10_LevInc	A11_LevEd	A12_LevEd	A13_Rank1	PaysVis	Mixit	Nblang	VecuEtranger
2	a lot of islands winding coastline small in area deve	WUH	HEA	F	2	Inc1	EduF2	EduM2	(4) Sup	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
3	a place for old-age pension	NKG	HEA	F	2	Inc1	EduF3	EduM3	(3) Nat	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
4	advanced civilized developed	CAN	BUS	M	2	Inc1	EduF2	EduM2	(3) Nat	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
5	advanced cultural sports environmental protection	CAN	HEA	M	2	Inc1	EduF1	EduM1	(3) Nat	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
6	advanced developed modern environmental prote	WUH	ART	M	1	Inc1	EduF2	EduM2	(3) Nat	Nb_Vis_O	Mixit_N	0-1lang	Vecu_etra_N
7	advanced flourishing dense population whitey	CAN	HEA	M	2	Inc1	EduF3	EduM3	(1) Loc	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
8	advanced science and technology	NKG	HEA	M	2	Inc1	EduF4	EduM4	(3) Nat	Nb_Vis_1a	Mixit_N	2lang	Vecu_etra_N
9	advanced welfare comfort life	SHA	ART	M	2	Inc1	EduF3	EduM2	(5) Glo	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
10	aging society developed many languages high stan	SHA	SHS	F	2	Inc1	EduF2	EduM3	(1) Loc	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
11	an organic whole	SHA	ENG	M	2	Inc1	EduF2	EduM3	(1) Loc	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
12	ancient peaceful developed open fervidly	WUH	HEA	F	2	Inc1	EduF1	EduM1	(1) Loc	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
13	art beautiful rich elegant power	CAN	SHS	F	3	Inc1	EduF2	EduM1	(1) Loc	Nb_Vis_O	Mixit_N	3lang	Vecu_etra_N
14	barbarous rich uncourteous advanced invasion	BJS	HEA	M	2	Inc1	EduF2	EduM2	(3) Nat	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N
15	beautiful abundant heaven distant shadowy	NKG	ENG	M	2	Inc1	EduF1	EduM1	(1) Loc	Nb_Vis_O	Mixit_N	2lang	Vecu_etra_N

(Source EuroBroadMap, 2009)

[2] Certains menus ou affichages peuvent être sensiblement différents selon la version utilisée.
[3] Le cas des textes longs est abordé à la fin de ce document



Importer un corpus

Comme ces textes très courts sont saisis dans un tableau, sélectionner *Tableur* et utiliser de préférence des fichiers « texte » avec délimiteur (.csv)[4].

Sélectionner la langue des textes du corpus (ici les réponses ont été saisies en anglais).

Découpage des textes (cas des fichiers textes bruts)

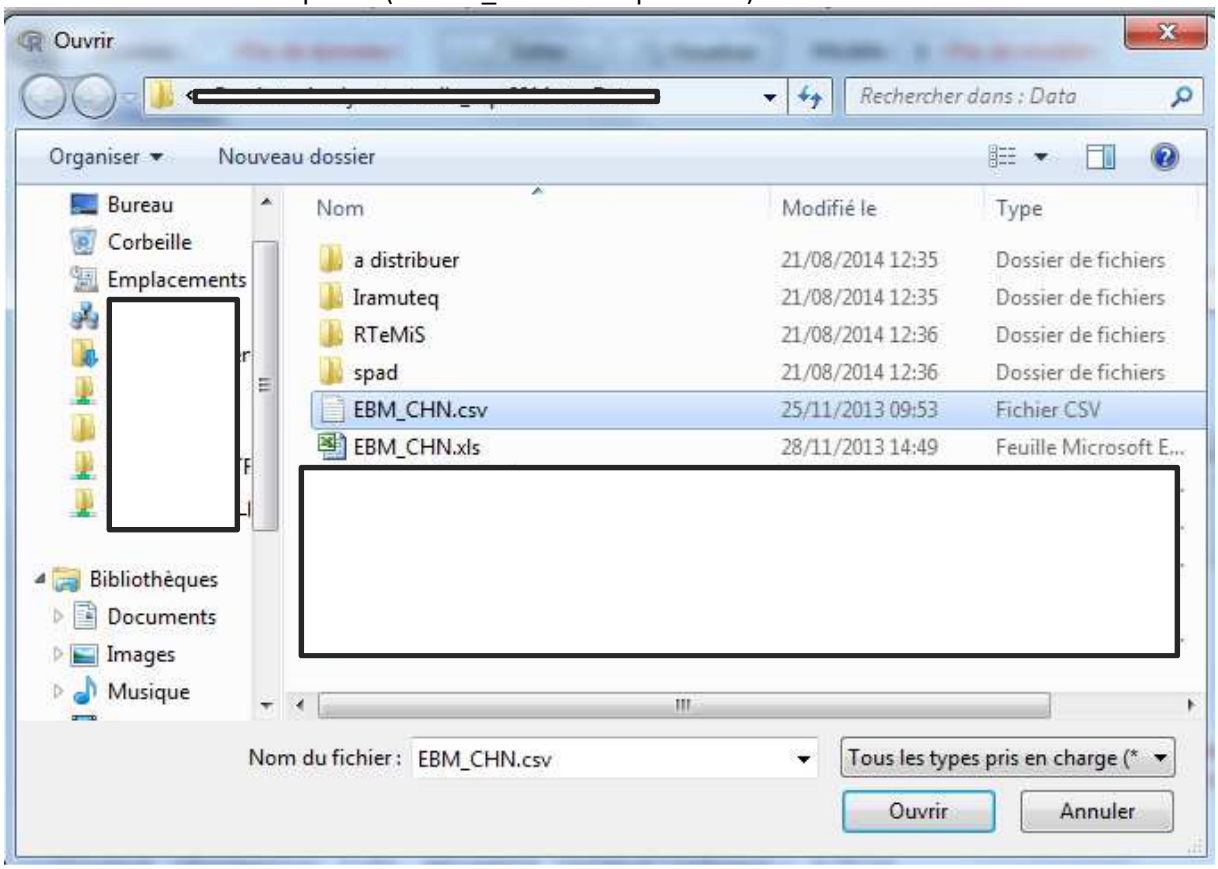
Si les textes sont longs, on peut les réduire en unités plus petites (appelées documents) par paragraphes. Le traitement de textes longs est abordé en fin de document.

[4]Les fichiers .xls ne sont pas toujours reconnus, selon la version.

Traitement des textes

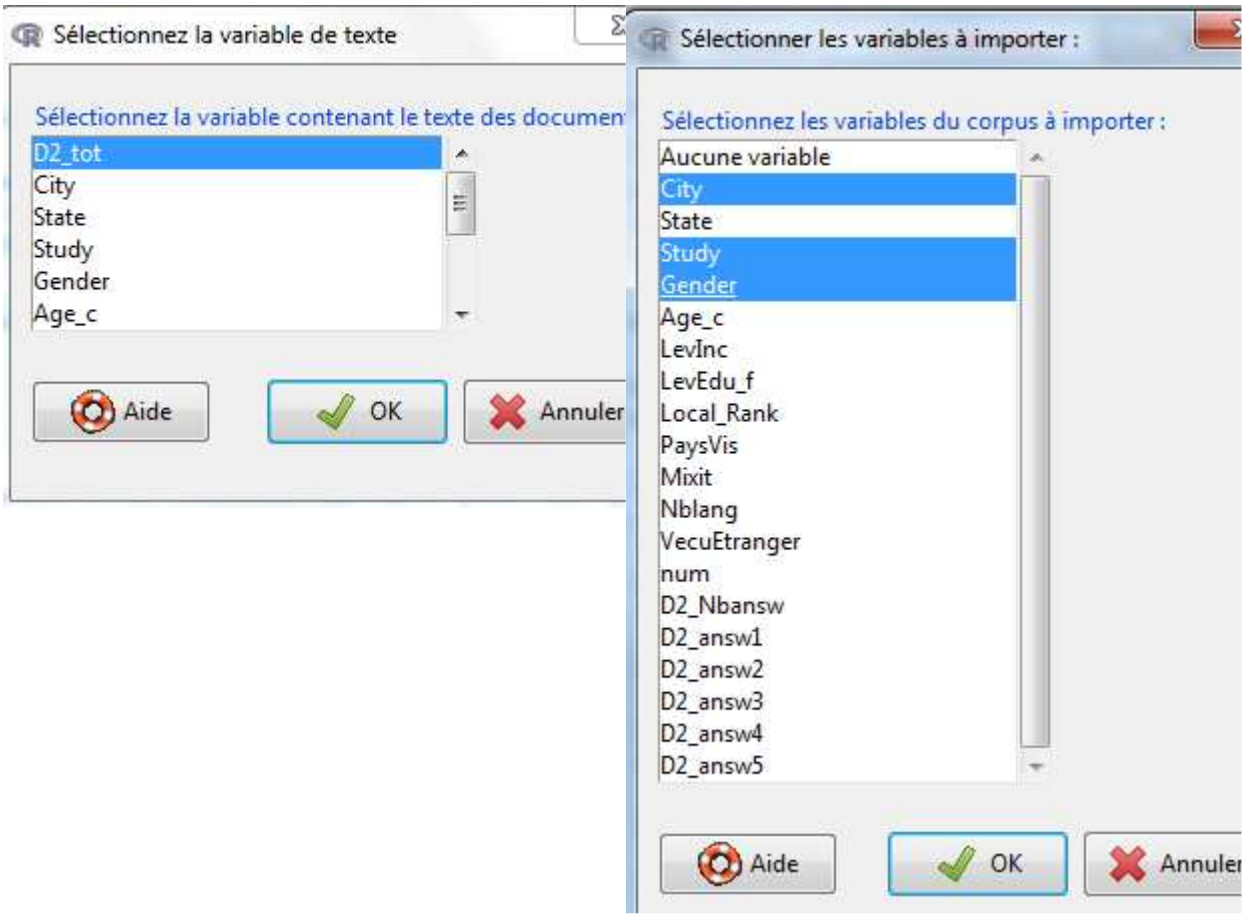
Si des mots ont été écrits en majuscules, ou que la première lettre d'un mot est en majuscule, cocher cette case *Ignorer la casse* permet d'éviter qu'une lettre écrite en majuscule rende un mot différent du même mot écrit tout en minuscules.
L'option *Extraire les radicaux* permet de regrouper sous le même terme les mots de même racine.
L'option *Editer la lemmatisation manuellement* permet de modifier la lemmatisation proposée par défaut (voir la fin du document).
Cocher *Supprimer les mots vides* correspond à supprimer les mot-outils (stopwords[5]) ; on peut aussi *Supprimer les nombres.*

Pointer sur le fichier à importer (Ici EBM_CHN.csv vu plus haut)



Sélectionner le fichier puis R.TeMiS demandera de préciser la variable texte à utiliser et de sélectionner les variables qualitatives à importer :

[5] Voir la liste des "Stopwords" de tm en français et en anglais en annexe.



Le corpus est chargé en mémoire.

```
> dtm
<<DocumentTermMatrix (documents: 1140, terms: 974)>>
Non-/sparse entries: 5010/1105350
Sparsity           : 100%
Maximal term length: 15
```

Le tableau lexical appelé dans R DocumentTermMatrix est formé de 1140 lignes (unités statistiques) et 974 colonnes (termes). Il contient 5010 cases des valeurs non nulles (occurrences) et 1105350 cases vides (on dit que c'est un tableau hyper-creux).
Les unités statistiques appelées « documents » correspondent aux 5 mots cités par les étudiants.
Le terme/mot le plus long est composé de 15 lettres.

Afficher le corpus actif

Permet de visualiser le corpus dans R.TeMiS.

(Source EuroBroadMap, 2009)

Ici, chaque réponse correspond à un document car il n'y a pas eu de découpage des textes en documents plus petits.

Afficher le lexique

Le menu Dictionnaire des termes affiche le lexique par ordre alphabétique (ou nombre d'occurrences).

Figure 3 : Affichage du lexique par ordre alphabétique (avec lemmatisation)

```
> attr(dict, "title") <- "Dictionnaire des termes par ordre alphabétique"

> dict
                Occurrences    Terme.Racine Occ. racine Mot vide Supprimé
abundance                 1           abund           8
abundant                  7           abund           8
abustle                   1          abustl           1
academic                  4          academ           4
accidental                1       accident           1
acquaintance              1       acquaint           1
acracholia                1     acracholia           1
active                    2           activ           2
admiring                  1           admir           1
advance                   1          advanc          97
advanced                 96          advanc          97
aegean                    1          aegean           1
aesthetics                1       aesthet           1
aestheticism              1     aesthetic           1
after                     1           after           1 Mot vide
against                   2         against           2 Mot vide
```

(Source EuroBroadMap, 2009)

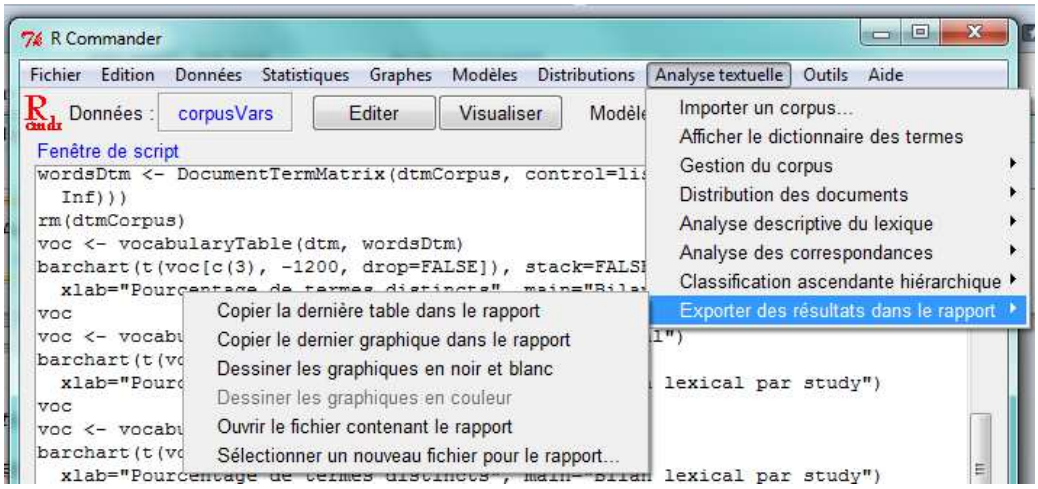
Lors de cette étape, on peut visualiser le résultat du traitement du corpus (extraction des radicaux et suppression des mots vides[6]).

[6] Pour visualiser la liste des mots vides en anglais (stopwords) lancer la commande stopwords("en") dans la fenêtre de script.

Encadré 2 : Sauvegarder les résultats au fur et à mesure des calculs

1. Générer un rapport

Les résultats des différentes opérations de l'analyse textuelle peuvent être sauvegardés au fur et à mesure dans un fichier par le menu *Exporter des résultats dans le rapport*. Cette procédure crée un fichier dans le navigateur (.html).



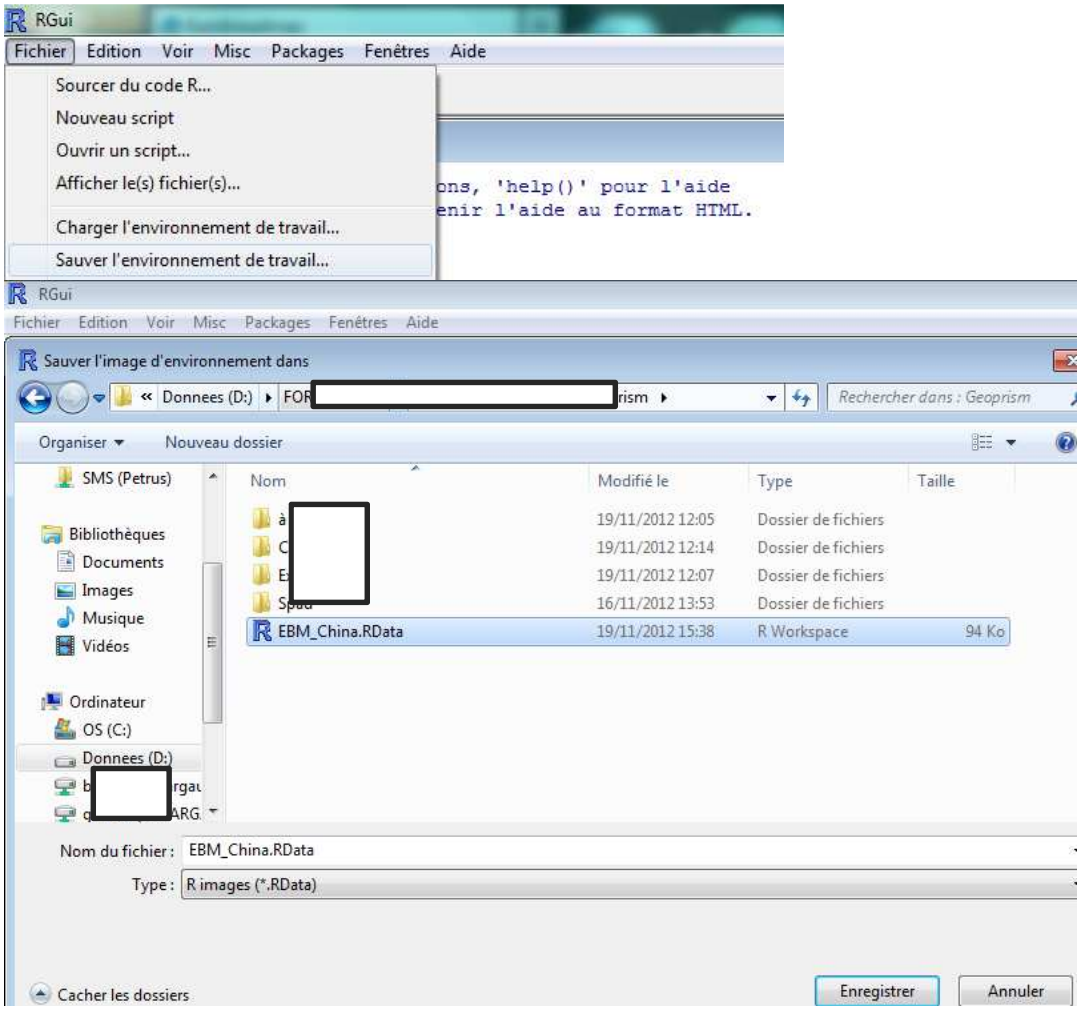
Une première fois aller sur le menu Sélectionner un nouveau fichier pour le rapport puis copier la dernière table ou le dernier graphique dans le rapport.

Pour visualiser chaque mise à jour cliquer sur Actualiser la page courante dans la barre Firefox.

Figure 4 : En tête du rapport généré par R.TeMiS

Sauvegarder uniquement le script généré par R Commander

A partir de la **fenêtre RGui**, il est possible de sauvegarder tous les résultats des calculs en cours d'utilisation (*Sauver l'environnement de travail*) pour pouvoir les réutiliser par la suite à l'ouverture d'une nouvelle session R (*Charger l'environnement de travail*).



2 Utilisation du Markdown

Permet de sauvegarder à la fois le script et les résultats dans un rapport. On se positionne sur l'onglet R Markdown puis on clique sur *Générer un rapport*

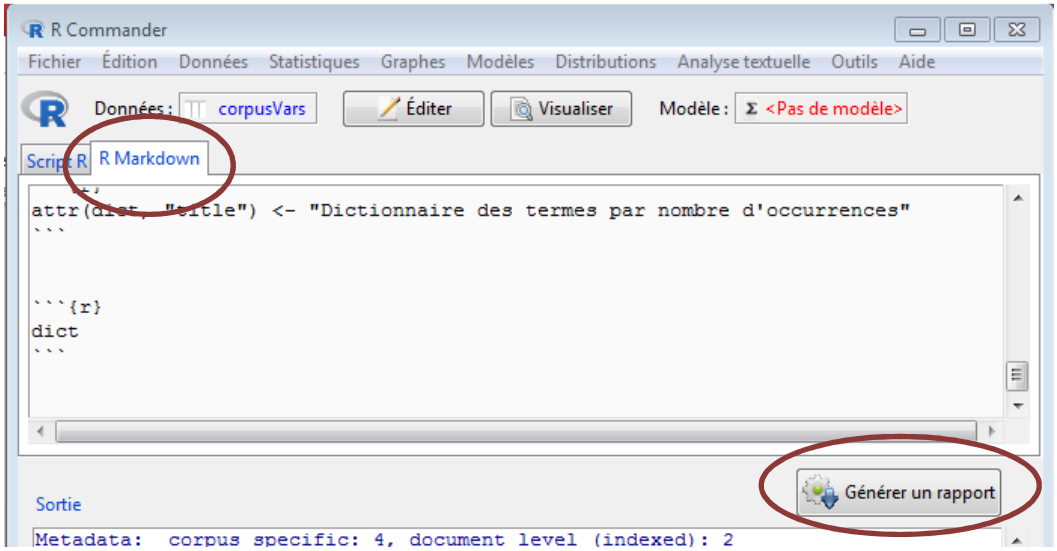


Figure 5 : Générer un rapport contenant le script et les résultats

```
> dict <- termsDictionary(dtm, "occurrences")
```

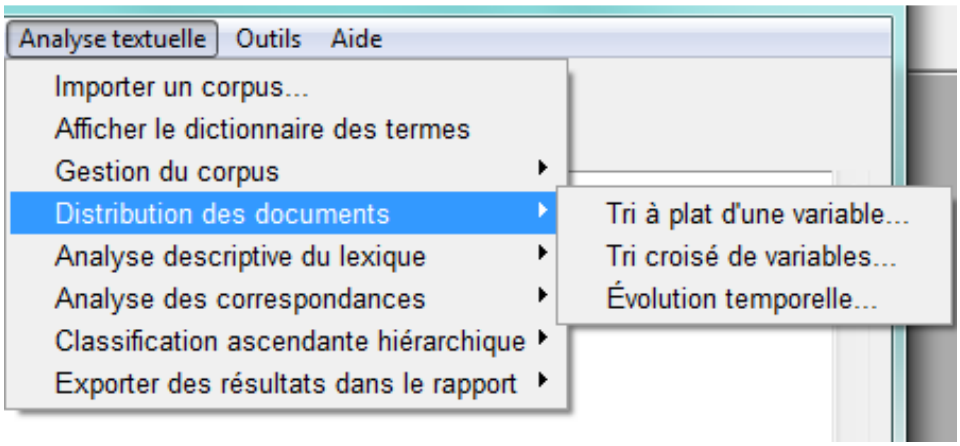
```
> attr(dict, "title") <- "Dictionnaire des termes par nombre d'occurrences"
```

```
> dict
```

```
                Occurrences Mot vide Supprimé
developed               355
rich                    244
romantic                204
beautiful               166
advanced                 96
civilized                78
of                       76 Mot vide Supprimé
and                      67 Mot vide Supprimé
small                    66
freedom                  65
open                     65
classical                64
high                     62
good                     59
elegant                  57
civilization             51
environment              47
clean                    41
welfare                  41
```

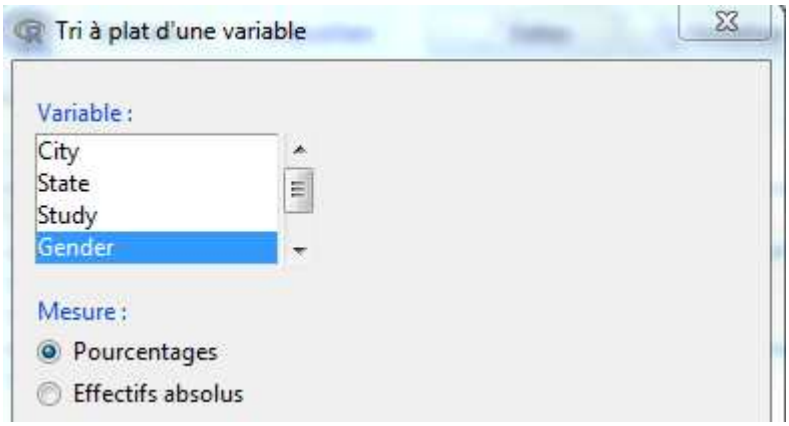
Décrire les métadonnées

Menu Distribution des documents



Permet de vérifier la distribution des variables que l'on veut utiliser dans les analyses.

Tri à plat de la variable Gender

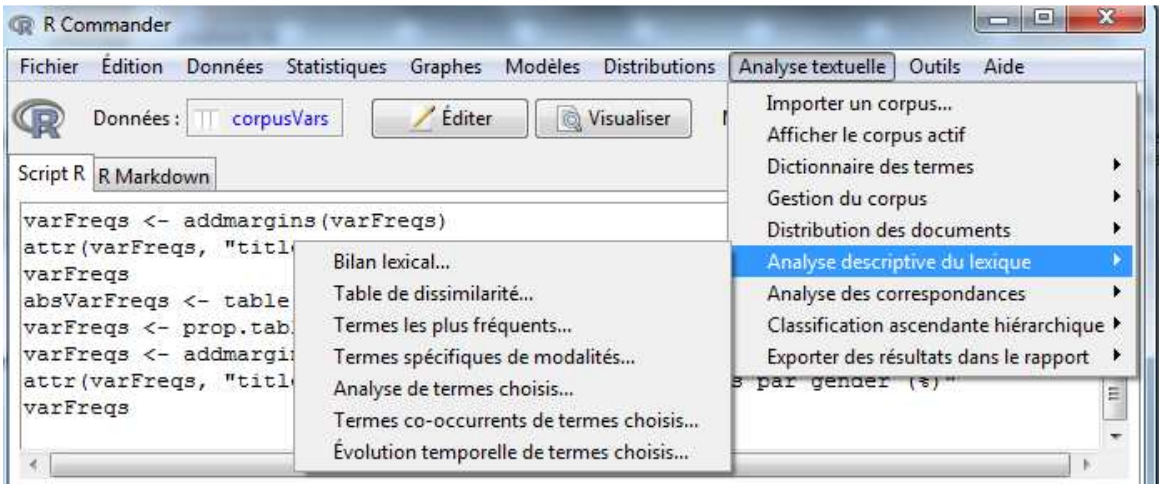



```
> varFreqs
Gender
  F   M Sum
 52  48 100
```

Dans ce corpus, la répartition des réponses (documents) entre femmes (F] et hommes (M) est assez équilibrée (52% et 48%).

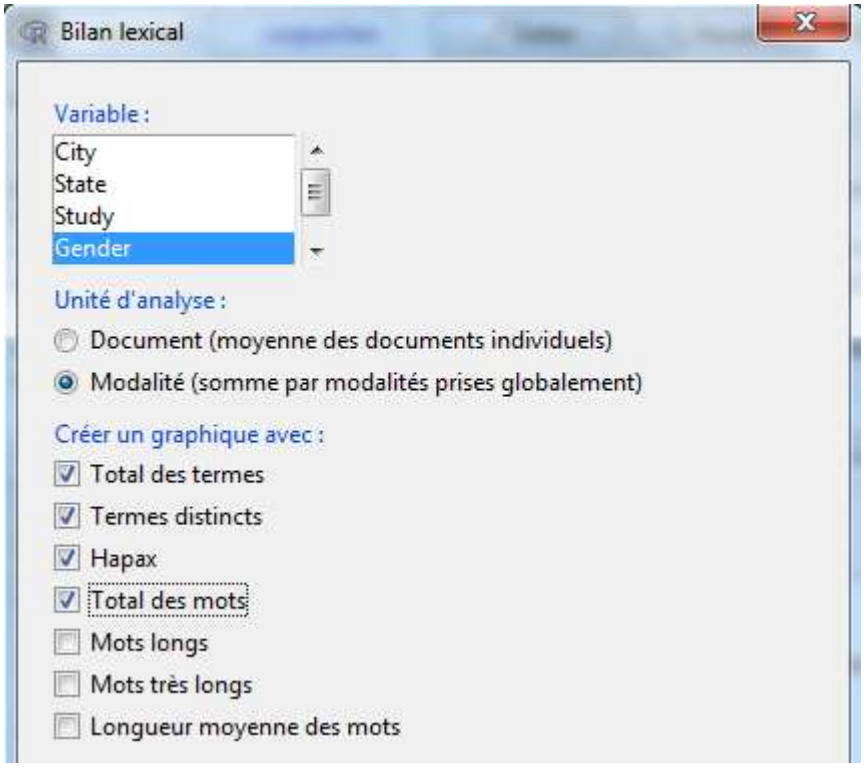
Décrire le corpus

Menu **Analyse descriptive du lexique**



Bilan lexical

Affiche, le nombre de termes, le nombre de termes distincts, le pourcentage de termes distincts par sous partie ou pour l'ensemble du corpus.



Sélectionner une variable et prendre la modalité comme unité d'analyse pour calculer un bilan lexical (ici sur les réponses des femmes et des hommes) et aussi visualiser le bilan global.

Il est aussi possible de générer le graphique associé.

```
> voc

Total par catégorie :                    F      M Total du corpus
  Nombre de termes                  2604.0 2455.0          5059.0
  Nombre de termes distincts         694.0  782.0          1134.0
  Pourcentage de termes distincts     26.7   31.9            22.4
  Nombre de hapax                    408.0  522.0           672.0
  Pourcentage de hapax                15.7   21.3            13.3
  Nombre de mots                    2604.0 2455.0          5059.0
  Nombre de mots longs              1649.0 1502.0          3151.0
  Pourcentage de mots longs           63.3   61.2            62.3
  Nombre de mots très longs          403.0  404.0           807.0
  Pourcentage de mots très longs      15.5   16.5            16.0
  Longueur moyenne des mots            7.2    7.1             7.2
```

(Source EuroBroadMap, 2009)

Sélectionner l'ensemble des documents permet de faire un bilan sur l'ensemble du corpus (ici les 1140 réponses).

On dénombre dans ce corpus 5059 mots dont 1134 mots différents[7]*. Les étudiantes citent moins de mots différents que les étudiants (694 contre 782).*

Sélectionner la variable Document si le corpus a été découpé en unités plus petites à l'importation (options Découpage du texte) (voir la partie importer un corpus à partir de textes « longs »).

[7] Cette mesure est un indicateur de « richesse » du vocabulaire

Table de dissimilarité

Calcule la distance du Khi2 entre documents (ici les réponses des étudiants à l'enquête) (si le nombre n'est pas trop grand) ou entre modalités d'une variable ou de deux variables qualitatives.


```
> diss
    ART BUS ENG HEA POL
BUS 1.8
ENG 1.9 1.9
HEA 1.8 1.7 1.8
POL 1.8 1.8 1.8 1.7
SHS 1.8 1.8 1.8 1.7 1.7
```

Le vocabulaire est plus proche entre les étudiants en sciences sociales (SHS), en santé (HEA) et sciences politiques (POL)

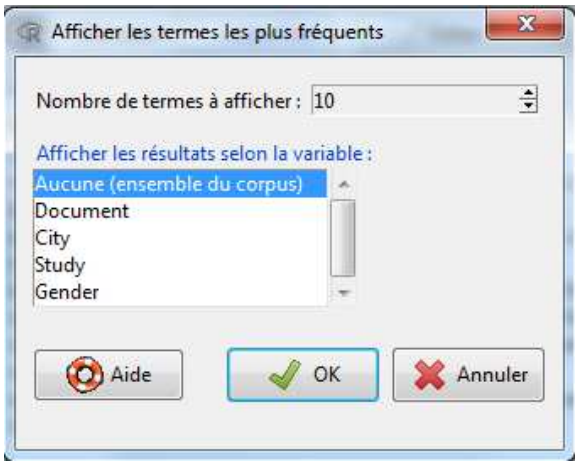

```
> diss
          F   M
ART 579 1.4 1.6
BUS 580 1.4 1.4
ENG 578 1.6 1.3
HEA 580 1.4 1.3
POL 578 1.2 1.3
SHS 575 1.1 1.4
```

Les mots cités par les filles sont plus proches des mots cités par les étudiants en sciences sociales (SHS) ou sciences politiques (POL)

(Source EuroBroadMap, 2009)

Termes les plus fréquents …

Affiche les termes les plus fréquents pour l'ensemble du corpus ou par modalité d'une variable qualitative donnée (métadonnées).



On peut choisir le nombre de termes à afficher (10 par défaut).

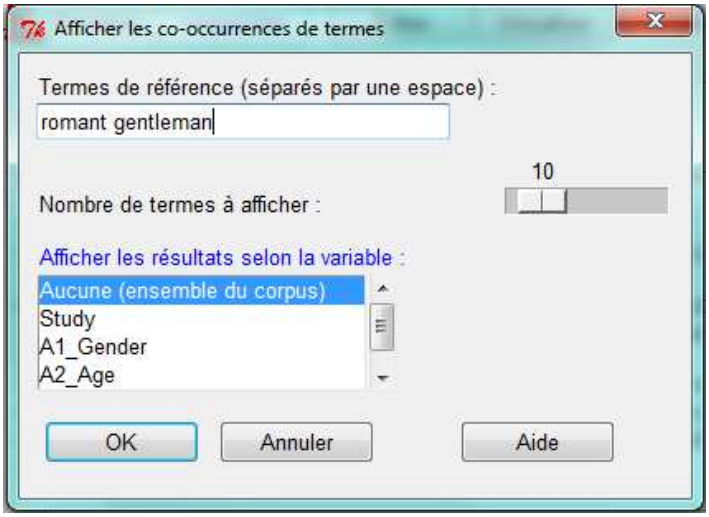
```
> freqTerms <- frequentTerms(dtm, NA, 10)

> attr(freqTerms, "title") <- "Termes les plus fréquents du corpus"

> freqTerms
        Occ. globales % global
develop           363      7.2
rich              245      4.8
romant            204      4.0
beauti            168      3.3
civil             130      2.6
advanc             97      1.9
of                 76      1.5
eleg               68      1.3
and                67      1.3
classic            67      1.3
attr(,"title")
[1] "Termes les plus fréquents du corpus"
```

(Source EuroBroadMap, 2009)

Rechercher des co-occurrences

Menu Analyse descriptive du lexique puis Termes co-occurrents de termes choisis…

Cherche les termes co-occurents à un ou plusieurs termes (mais pas entre ces termes) pour l'ensemble du corpus ou par sous partie.



Possibilité de travailler par sous corpus selon les modalités d'une variable.

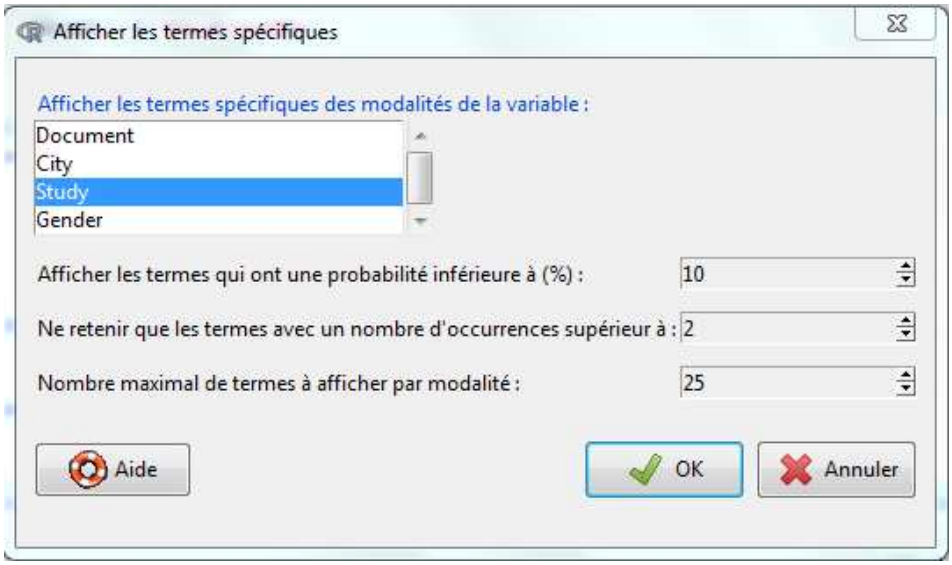
```
> coocs <- sapply(c("romant", "gentleman"), termChisqDist, dtm, 10, simplify=FALSE)

> coocs
$romant
     prolif       layer    champion    unreason      manorc        nato      leader      stilli
   6.240318    6.265567    6.279783    6.495112    6.640877    6.646129    6.698307    6.725162
   contamin flatterrain
   6.772554    6.826062

$gentleman
     slouch       prolif       layer    champion    unreason        nato      leader      stilli
   170.1637    175.2150    175.2255    175.2397    175.4550    175.6061    175.6582    175.6851
   contamin flatterrain
   175.7325    175.7860
```

(Source EuroBroadMap, 2009)

Calculer des spécificités

Menu Analyse descriptive du lexique puis Termes spécifiques de modalités …

Permet de repérer le vocabulaire spécifique ou « sur employé » pour chaque modalité d'une variable qualitative choisie.



On peut réduire le lexique en ne gardant que les termes de fréquence supérieure à un seuil (5 par défaut).

```
> attr(specTerms, "title") <- "Termes spécifiques par Study"

> specTerms
$ART
           % terme/mod. % mod./terme % global Modalité Global Valeur t Proba.
pretty             1.68         80.0    0.296       12     15       5.5 0.0000
comfort            0.84        100.0    0.119        6      6       4.3 0.0000
fine               1.26         69.2    0.257        9     13       4.3 0.0000
fashionable        1.54         52.4    0.415       11     21       4.0 0.0000
art                1.68         46.2    0.514       12     26       3.8 0.0001
prosperous         1.68         38.7    0.612       12     31       3.2 0.0006
free               0.70         71.4    0.138        5      7       3.1 0.0009
leisure            1.68         33.3    0.711       12     36       2.8 0.0027
arts               0.56         66.7    0.119        4      6       2.6 0.0046
old                0.98         38.9    0.356        7     18       2.4 0.0083
grade              0.42         75.0    0.079        3      4       2.3 0.0100
easy               0.56         50.0    0.158        4      8       2.1 0.0171
design             0.28        100.0    0.040        2      2       2.1 0.0199
```

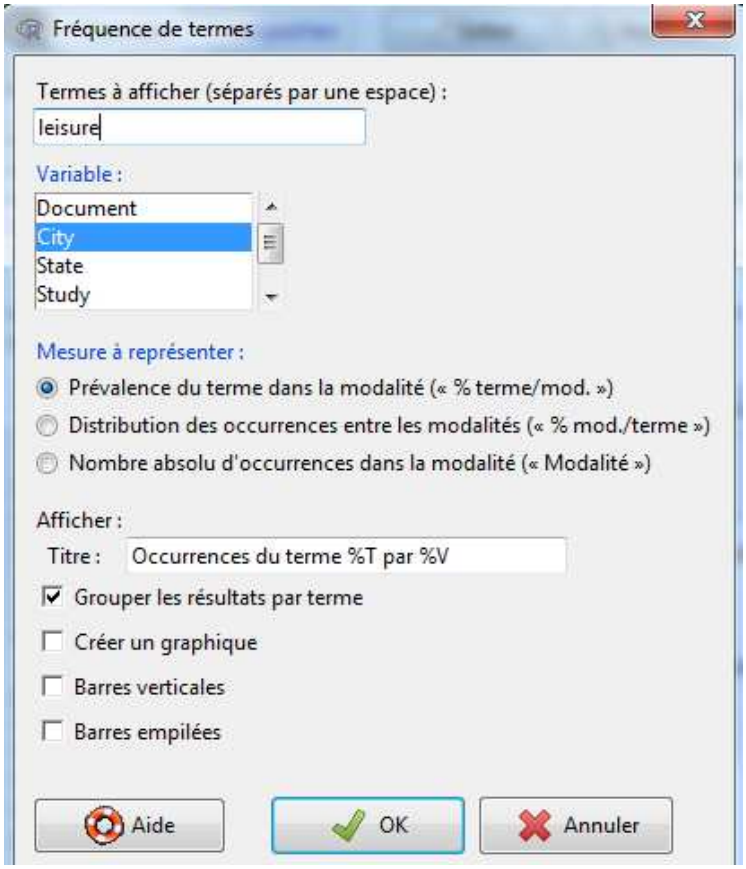
(Source EuroBroadMap, 2009)

Les termes sont triés par valeur test, les valeurs les plus positives en haut et les plus négatives en bas. Une valeur test positive indique que le mot est sur représenté (on peut dire aussi sur employé) dans la catégorie (ici Study), une valeur négative indique qu'il est sous représenté.

Le mot/terme pretty représente 1,68% de l'ensemble des occurrences des mots cités par les étudiants en ART. 80% des occurrences du mot pretty sont citées par des étudiants en ART. Le mot pretty est très spécifique aux étudiants en ART

Fréquence de termes

On choisit un ou plusieurs termes[8] pour lesquels on veut connaître leur fréquence (ici *leisure*) dans le corpus et aussi leur spécificité.

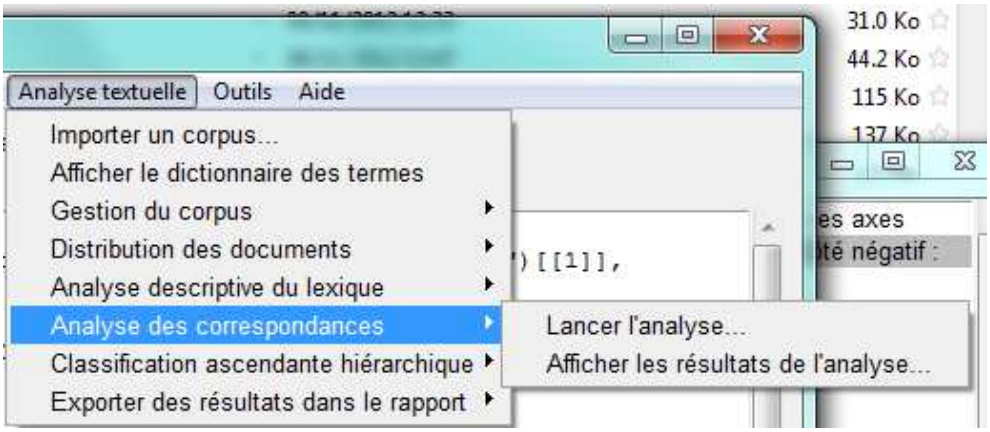



```
> termFreqs
, , leisure

    % terme/mod. % mod./terme % global Modalité Global Valeur t Proba.
BJS         0.64           17     0.71        6     36  -0.0095   0.50
CAN         0.94           28     0.71        9     36   0.8145   0.21
NKG         0.72           19     0.71        6     36  -0.1289   0.55
SHA         0.72           22     0.71        7     36  -0.1361   0.55
WUH         0.51           14     0.71        5     36  -0.5998   0.27

attr(,"title")
[1] "Occurrences du terme leisure par city (en % de tous les termes)"
```

Le mot leisure a été cité 36 fois. Le terme/mot leisure représente 0,94% des mots cités par les étudiants de Canton (CAN). 28% des occurrences du mot leisure sont données par les étudiants de Canton.

[8] Si on a coché *Extraire les radicaux* à l'importation, il faudra indiquer la forme racine (lemme).

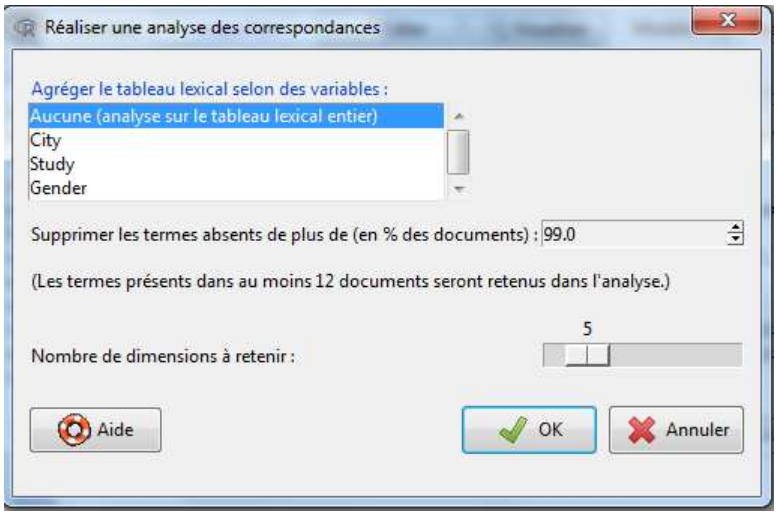
Menu Analyse des correspondances (AFC)

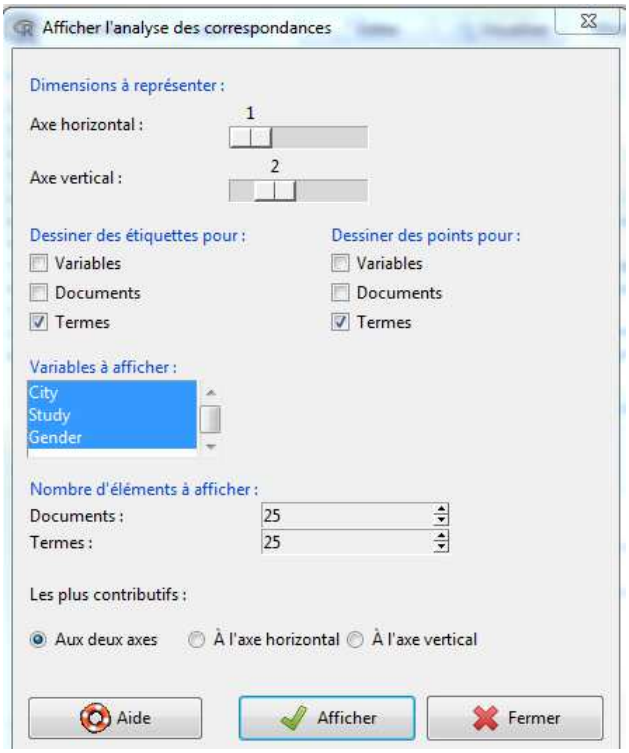


AFC sur un tableau lexical entier (TLE)

L'analyse factorielle des correspondances sur le tableau lexical entier (TLE) met en évidence des mots co-occurrents et les représente sur des graphiques appelés plans factoriels. La lecture des mots les plus contributifs aux axes et des concordances permet d'identifier des champs lexicaux ou thématiques.

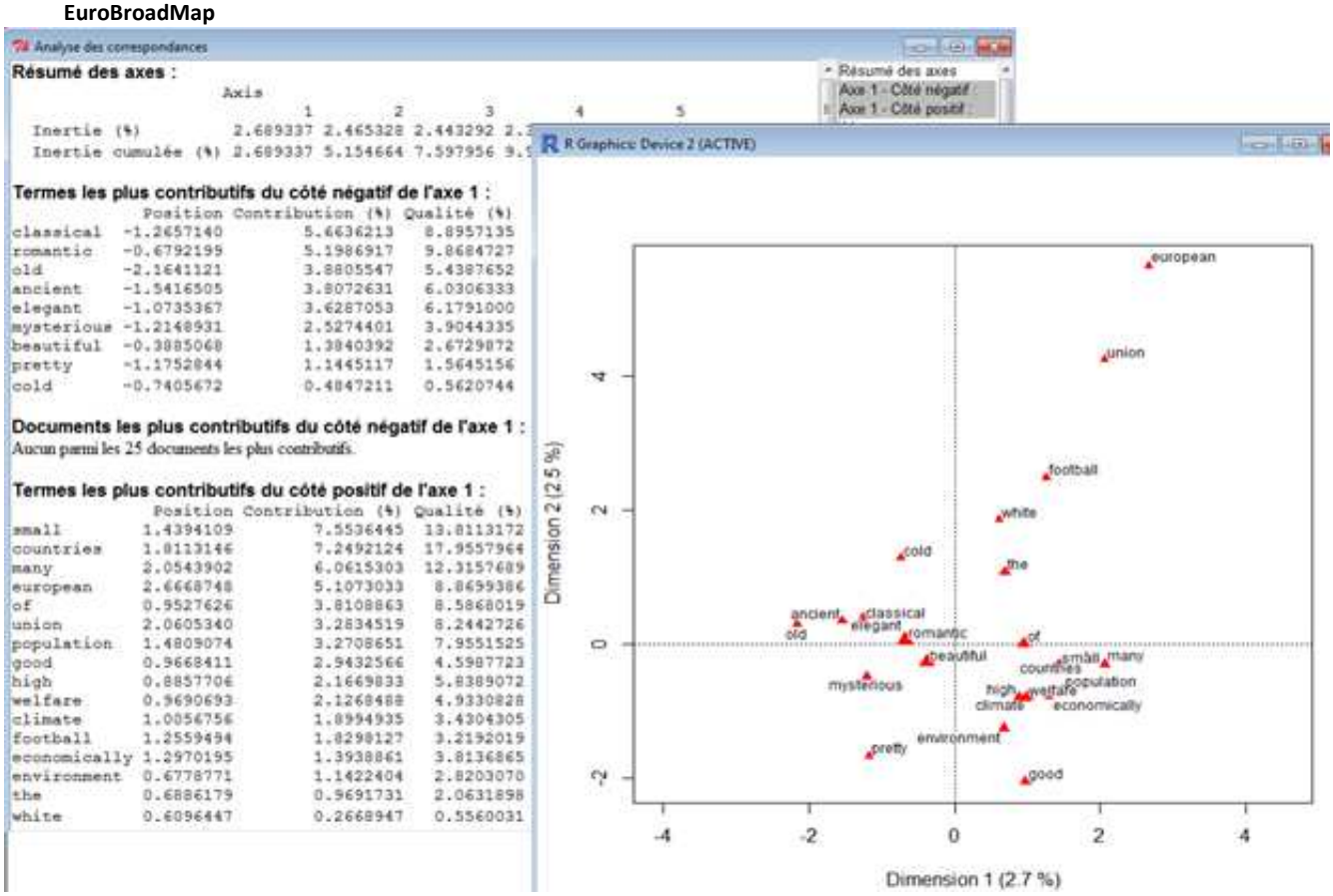
Dans la pratique, on sélectionne Aucune (analyse sur le tableau lexical entier) c'est-à-dire que l'on ne va pas agréger le tableau lexical selon des variables qualitatives.





Ici les variables (catégories) sont des éléments illustratifs. On peut choisir d'afficher les termes les plus contributifs (par plan ou axe par axe) et leur nombre.

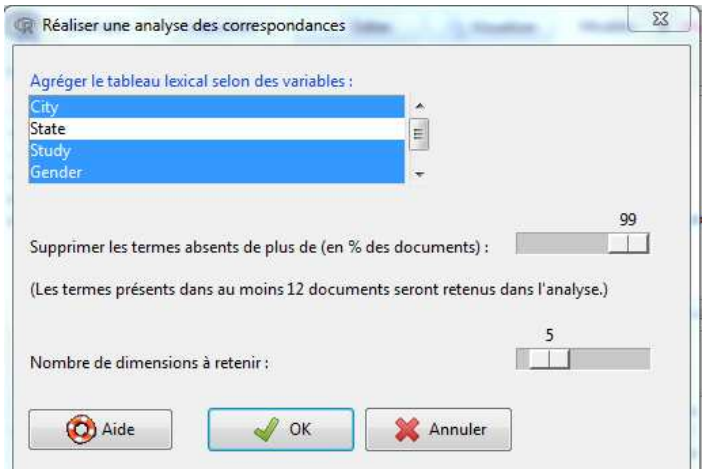
Figure 6 : AFC sur le TLE - Mots cités par les étudiants interrogés en Chine - Plan 1-2 et éléments contributifs EuroBroadMap



(Source EuroBroadMap, 2009)

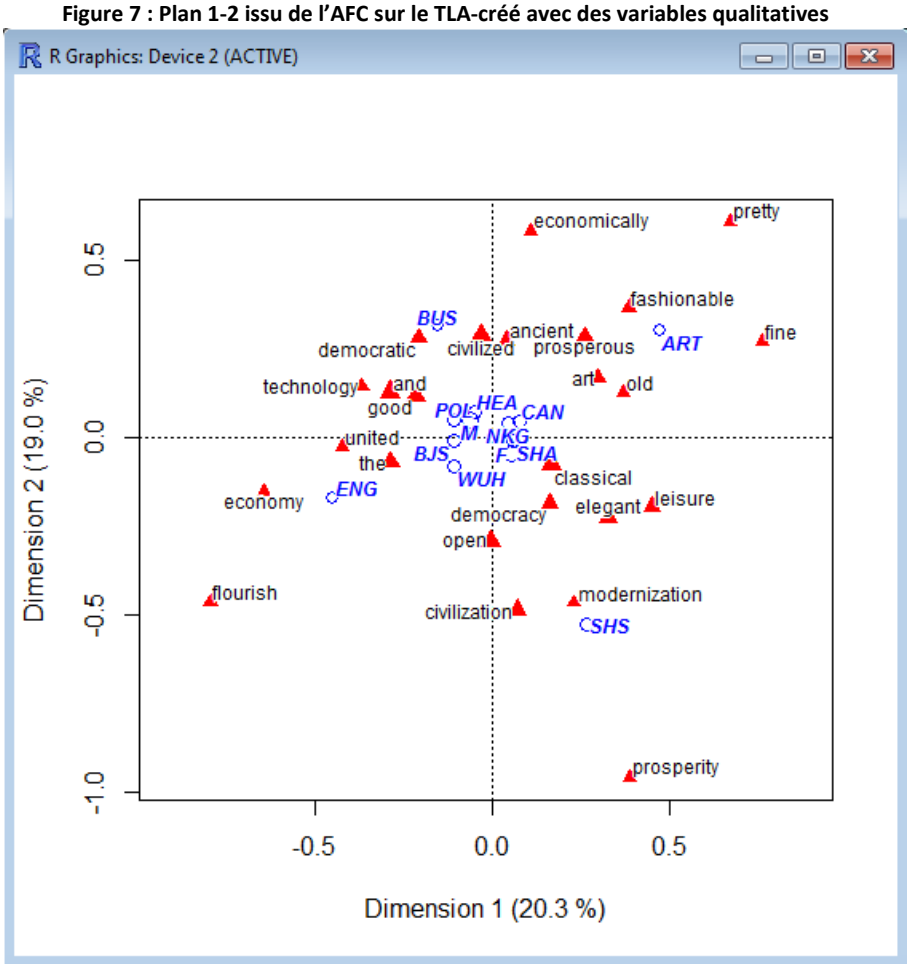
AFC sur un tableau lexical agrégé (TLA)

On choisit les variables (ici les caractéristiques des répondants) pour créer le tableau de contingence (croisant les mots du corpus et les modalités des variables qualitatives).



Ici les modalités des variables contribuent aux axes (elles sont « actives »).

Figure 7 : Plan 1-2 issu de l'AFC sur le TLA-créé avec des variables qualitatives



(Source EuroBroadMap, 2009)

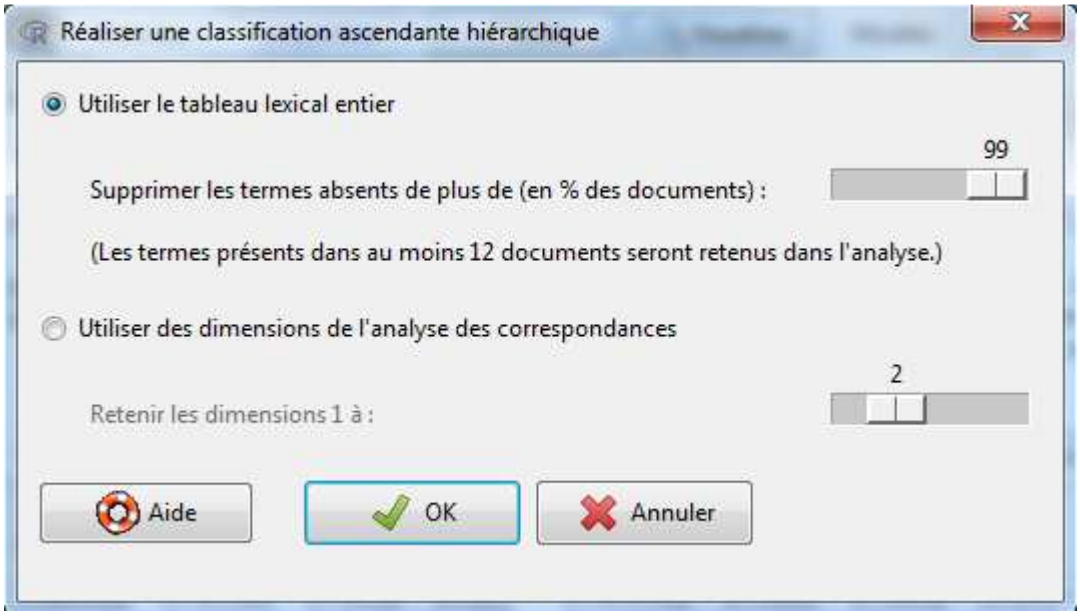
Effectuer une AFC sur le TLA permet de structurer l'ensemble des « mots » en fonction des caractéristiques des étudiants (ici ville, domaine d'étude et genre) et de répondre à la question *Qui dit quoi ?* On peut interpréter plus finement les proximités graphiques entre les mots et les caractéristiques individuelles en recourant au calcul des *termes spécifiques* (vu plus haut).

Classification ascendante hiérarchique

Sur le TLE ou sur les facteurs de l'AFC sur le TLA ou sur les facteurs de l'AFC sur le TLE

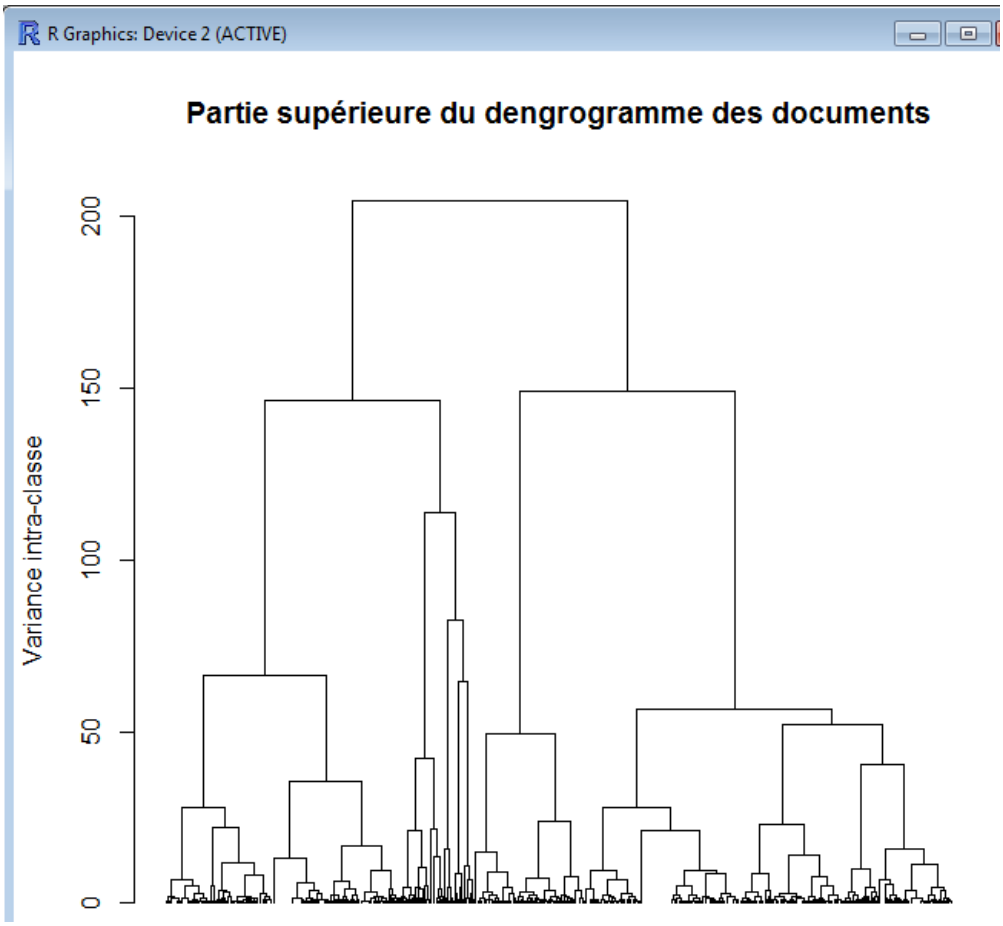
En classant ensemble des textes contenant des mots co-occurrents, cette méthode permet de repérer les thématiques du corpus.

La lecture des termes et des documents spécifiques de chaque classe « aide » à leur donner un intitulé.

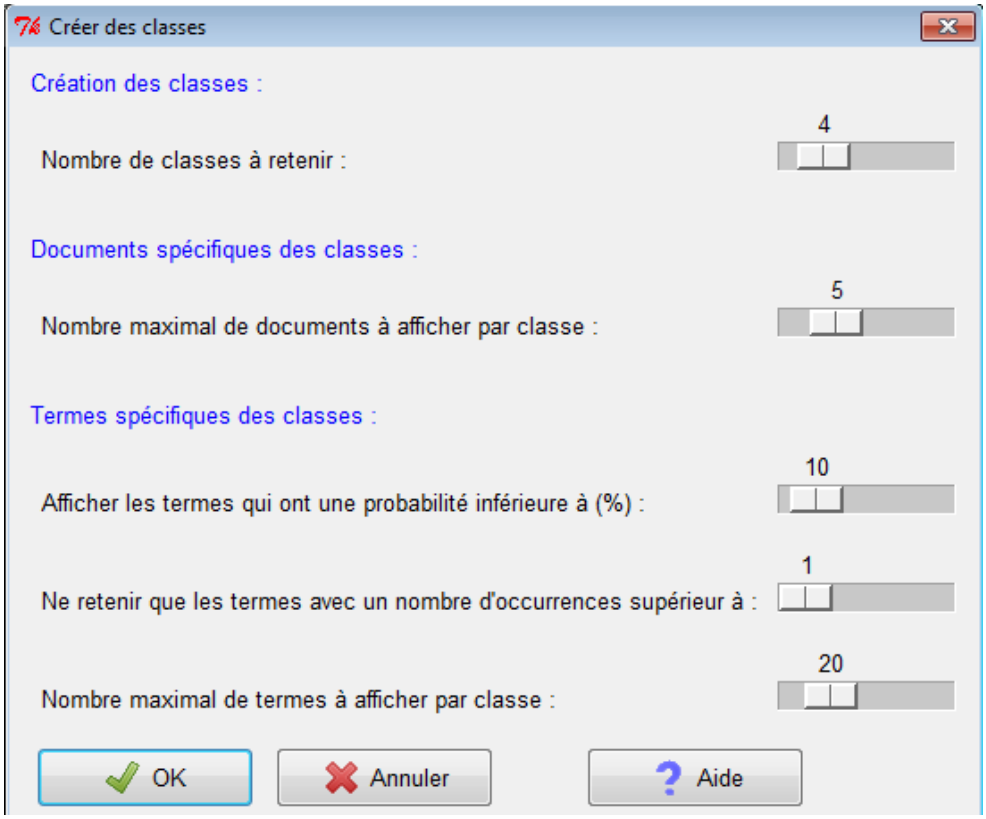


On peut aussi exécuter la classification sur les axes de la dernière AFC réalisée.

Permet de créer une nouvelle variable (métadonnée) que l'on pourra projeter sur le plan factoriel d'une AFC sur le TLA



On s'aide de dendrogramme pour déterminer le nombre de classes à retenir



Pour aller plus loin ….

Au vu des premiers résultats, il est souvent nécessaire de refaire des analyses complémentaires comme la recherche de concordances, de modifier le corpus analysé (création d'un sous corpus, suppressions de mots) ou d'intervenir dans la lemmatisation qui a été faite automatiquement dans un premier temps.

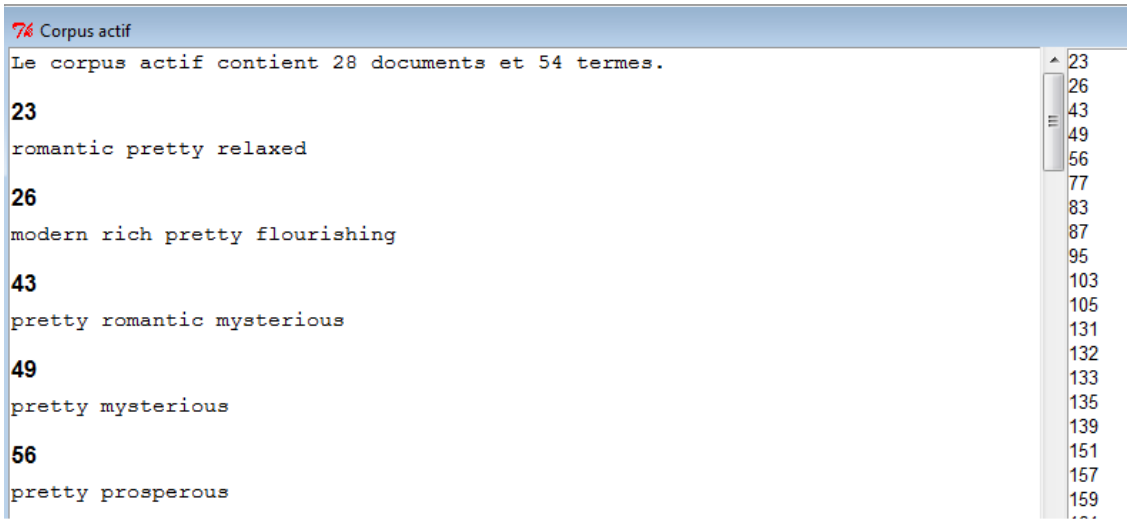
Nous décrivons ces opérations dans la partie suivante :

Afficher des concordances

Lire le contexte d'utilisation de certains mots (*concordances*) est une aide à l'interprétation complémentaire et indispensable à l'analyse des résultats tels que les mots spécifiques, plans factoriels ou classifications.

Cette étape consiste à restituer les réponses ou parties de textes dans lesquelles un terme donné est utilisé et donc à restreindre le corpus. Dans l'exemple ci-dessous, on a retenu les réponses contenant le terme *pretty*:

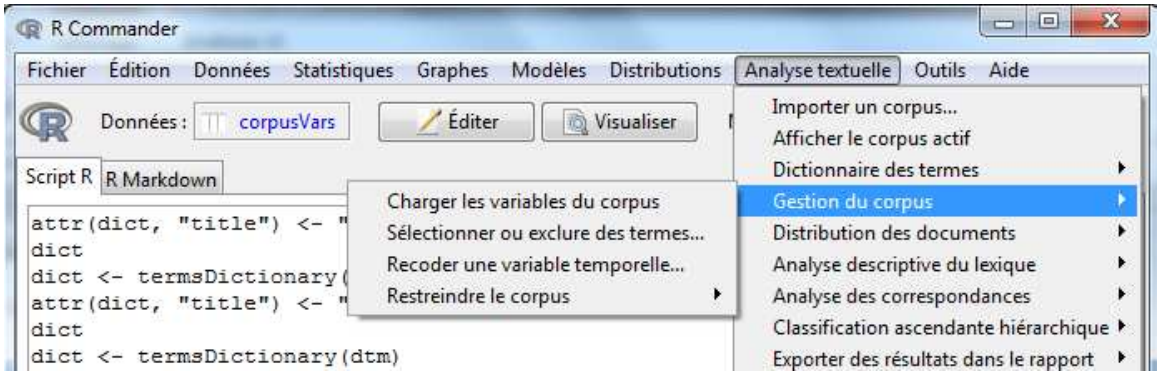
Figure 8 : Extrait des concordances du mot *pretty* dans les réponses des étudiants chinois



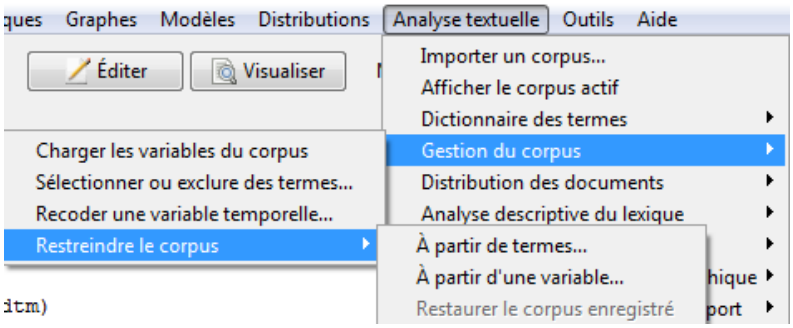
(Source EuroBroadMap, 2009)

Cette étape est décrite dans la partie Gestion du corpus.

Gestion du corpus



Il s'agit de sélectionner une sous partie du corpus à analyser en fonction des modalités d'une variable qualitative (métadonnée) ou de sélectionner (ou exclure) des textes contenant des termes spécifiés.



Deux types des conditions sont possibles dans la sélection de sous corpus :

Figure 9 : Condition sur les mots pour la sélection du sous corpus

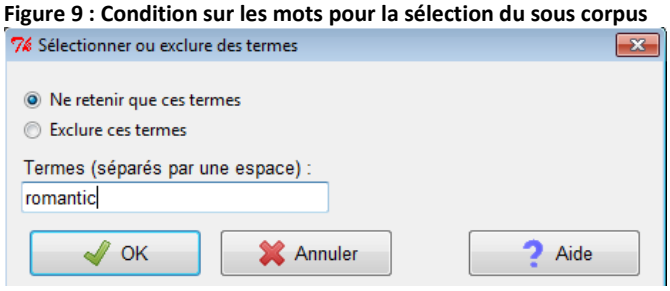
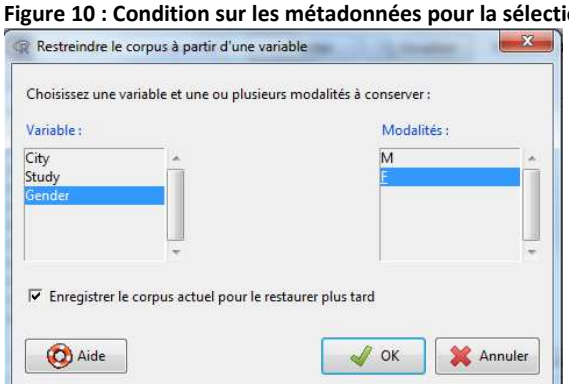


Figure 10 : Condition sur les métadonnées pour la sélection du sous corpus

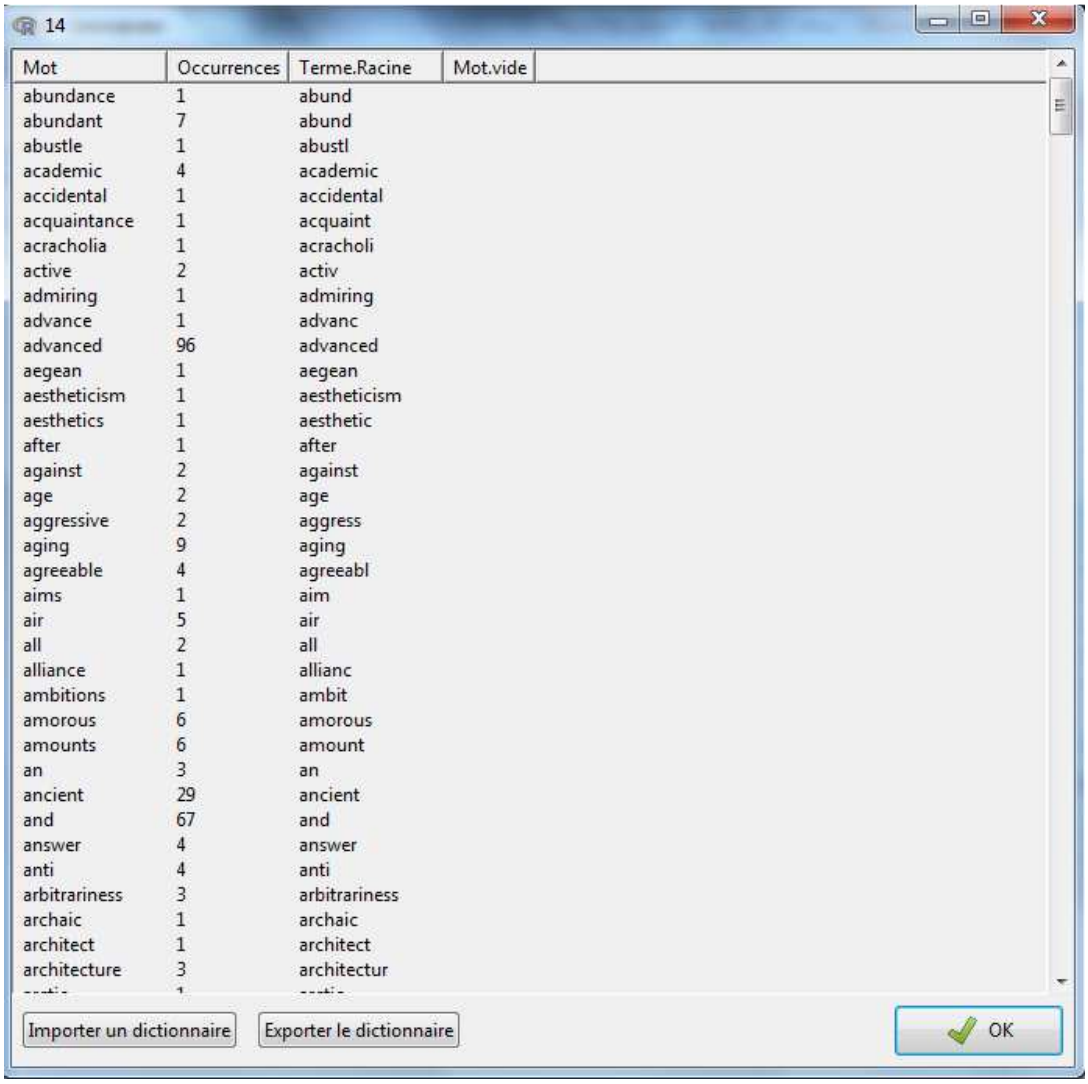


Si besoin, on pourra restaurer le corpus initial.

Modifier la lemmatisation

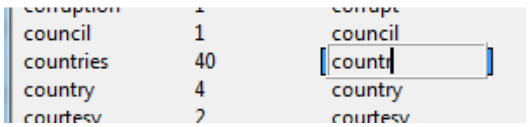
Il est possible de modifier le lexique issu de la lemmatisation automatique (option Extraire les radicaux (lemmatisation)) à l'importation du corpus. Pour cela, dans le menu d'importation et il faut cocher la case Editer la lemmatisation[9].

Une fenêtre de paramétrage permet d'intervenir sur les regroupements ou de corriger les Termes.



Mot	Occurrences	Terme.Racine	Mot.vide
abundance	1	abund	
abundant	7	abund	
abustle	1	abustl	
academic	4	academic	
accidental	1	accidental	
acquaintance	1	acquaint	
acracholia	1	acracholi	
active	2	activ	
admiring	1	admiring	
advance	1	advanc	
advanced	96	advanced	
aegean	1	aegean	
aestheticism	1	aestheticism	
aesthetics	1	aesthetic	
after	1	after	
against	2	against	
age	2	age	
aggressive	2	aggress	
aging	9	aging	
agreeable	4	agreeabl	
aims	1	aim	
air	5	air	
all	2	all	
alliance	1	allianc	
ambitions	1	ambit	
amorous	6	amorous	
amounts	6	amount	
an	3	an	
ancient	29	ancient	
and	67	and	
answer	4	answer	
anti	4	anti	
arbitrariness	3	arbitrariness	
archaic	1	archaic	
architect	1	architect	
architecture	3	architectur	

Il est possible de modifier le regroupement « automatique » des mots ou de changer le terme racine.



[9] Selon la version de R.TeMIS, charger le package methods avant l'importation avec la commande >library (methods)

Importer un corpus constitué de textes « longs »

Dans l'exemple des réponses des étudiants en Chine à la question : « *Citez 5 mots …* » nous pouvons considérer que les **documents sont « courts »** et qu'ils peuvent être saisis dans un fichier type « tableur » contenant en lignes les réponses et en colonnes différentes variables qualitatives comme le sexe, la ville d'enquête, …. Au paramétrage de l'importation du corpus, on a donc spécifié que le fichier était de type **tableur**.

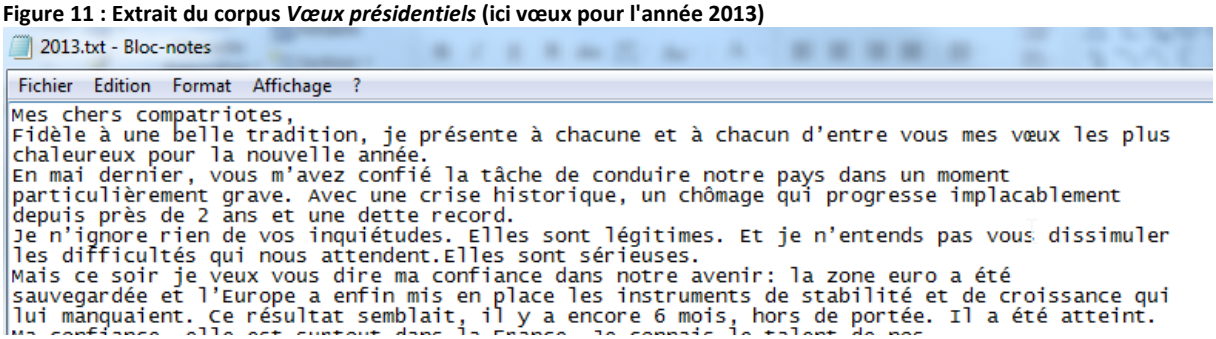
Si le corpus correspond à un ensemble de documents plus **« longs »** comme des chapitres d'ouvrages, des entretiens, des articles … et que l'on possède aussi des caractéristiques sur ces corpus, il faut **organiser les fichiers** comme suit :

Les différents documents constituant le corpus sont des fichiers texte (.txt) et doivent être placés dans un seul **répertoire** ne contenant que ces fichiers.

Les caractéristiques (métadonnées) sur ces documents peuvent être saisies dans un fichier de type « **tableur** » mais il faudra respecter l'ordre des lignes du tableau qui doit correspondre à l'ordre des différents documents dans le répertoire.

Ici par exemple on veut analyser l'ensemble des 5 vœux du président de la République aux Français (variable textuelle) et on a pour ces 5 textes, le nom du président (ici François Hollande) et les années concernées (2013, 2014, 2015, 2016, 2017).

Figure 11 : Extrait du corpus *Vœux présidentiels* (ici vœux pour l'année 2013)



Source http://www.elysee.fr/declarations/article/v-ux-du-president-de-la-republique-aux-francais/

Contenu du répertoire contenant les textes

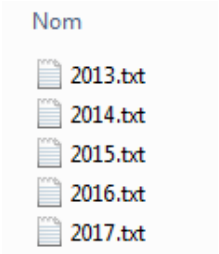


Tableau contenant les caractéristiques sur les textes

num	an	nom
1	2013	Hollande
2	2014	Hollande
3	2015	Hollande
4	2016	Hollande
5	2017	Hollande

Au moment du paramétrage de l'importation du corpus, il faut spécifier le nom répertoire contenant les fichiers bruts puis paramétrer le découpage des textes (càd cocher la case) et choisir le nombre de paragraphes qui correspondra au critère de découpage :

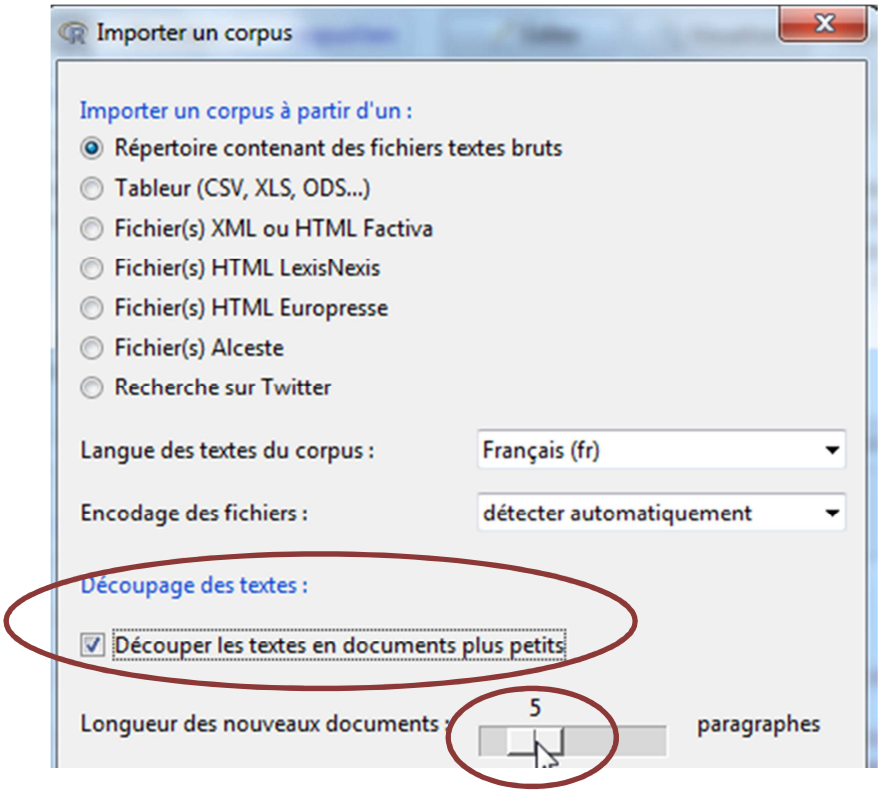
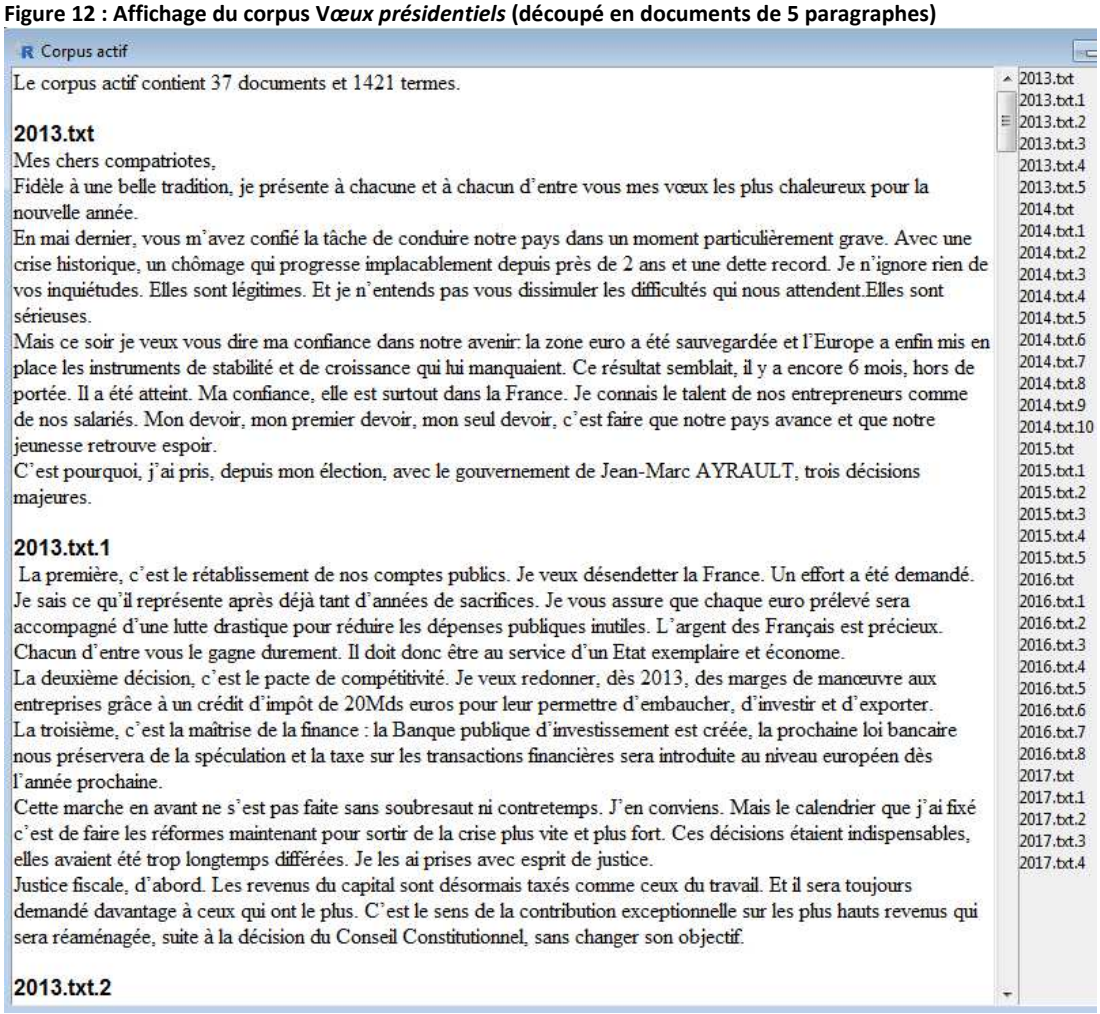
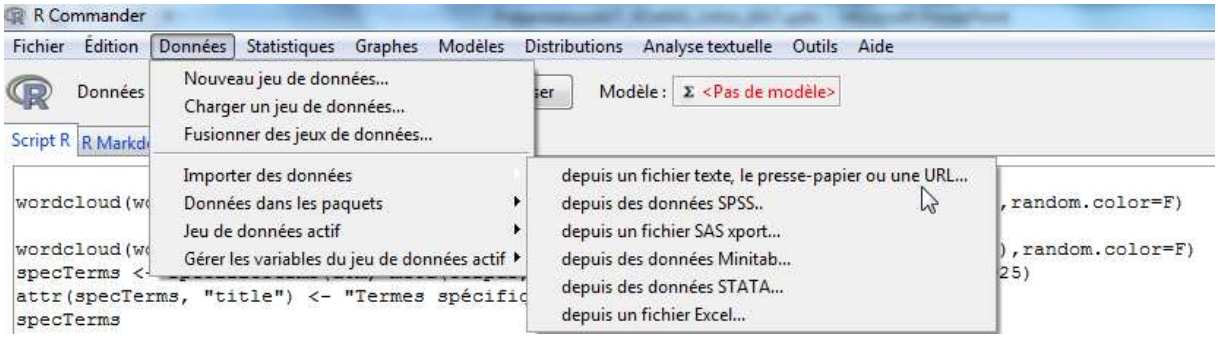
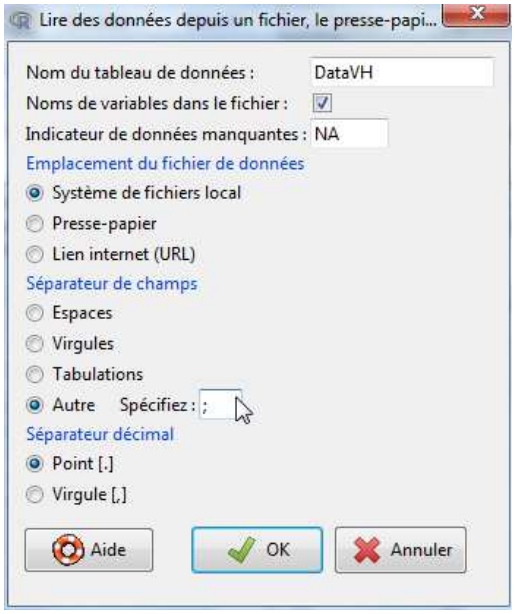


Figure 12 : Affichage du corpus V*œux présidentiels* (découpé en documents de 5 paragraphes)

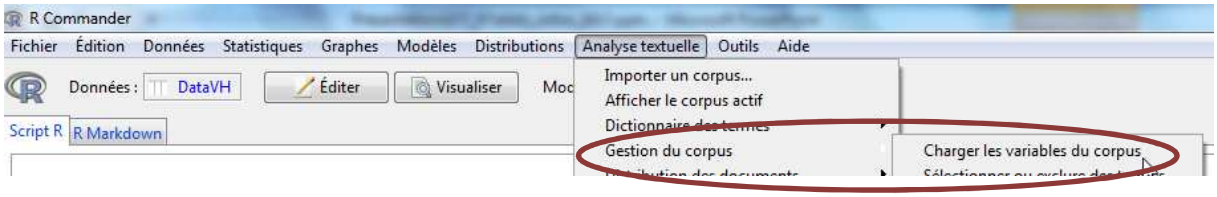


Pour utiliser les caractéristiques des textes, il faudra *importer les données* avec le menu Données de **R Commander**





Les données (ici data VH) seront chargées dans RCommander, il reste à les charger dans R.TeMiS pour les utiliser avec le corpus.



On pourra ensuite utiliser les méthodes de la statistique textuelle décrites dans le document.

Annexe : Lemmatisation de tm : liste des Stopwords

Figure 13 : Liste10 des "Stopwords" de tm (fr)

```
> stopwords("fr")
  [1] "au"       "aux"      "avec"     "ce"       "ces"      "dans"     "de"       "des"      "du"       "elle"
 [11] "en"       "et"       "eux"      "il"       "je"       "la"       "le"       "leur"     "lui"      "ma"
 [21] "mais"     "me"       "même"     "mes"      "moi"      "mon"      "ne"       "nos"      "notre"    "nous"
 [31] "on"       "ou"       "par"      "pas"      "pour"     "qu"       "que"      "qui"      "sa"       "se"
 [41] "ses"      "son"      "sur"      "ta"       "te"       "tes"      "toi"      "ton"      "tu"       "un"
 [51] "une"      "vos"      "votre"    "vous"     "c"        "d"        "j"        "l"        "à"        "m"
 [61] "n"        "s"        "t"        "y"        "été"      "étée"     "étées"    "étés"     "étant"    "suis"
 [71] "es"       "est"      "sommes"   "êtes"     "sont"     "serai"    "seras"    "sera"     "serons"   "serez"
 [81] "seront"   "serais"   "serait"   "serions"  "seriez"   "seraient" "étais"    "était"    "étions"   "étiez"
 [91] "étaient"  "fus"      "fut"      "fûmes"    "fûtes"    "furent"   "sois"     "soit"     "soyons"   "soyez"
[101] "soient"   "fusse"    "fusses"   "fût"      "fussions" "fussiez"  "fussent"  "ayant"    "eu"       "eue"
[111] "eues"     "eus"      "ai"       "as"       "avons"    "avez"     "ont"      "aurai"    "auras"    "aura"
[121] "aurons"   "aurez"    "auront"   "aurais"   "aurait"   "aurions"  "auriez"   "auraient" "avais"    "avait"
[131] "avions"   "aviez"    "avaient"  "eut"      "eûmes"    "eûtes"    "eurent"   "aie"      "aies"     "ait"
[141] "ayons"    "ayez"     "aient"    "eusse"    "eusses"   "eût"      "eussions" "eussiez"  "eussent"  "ceci"
[151] "cela"     "celà"     "cet"      "cette"    "ici"      "ils"      "les"      "leurs"    "quel"     "quels"
[161] "quelle"   "quelles"  "sans"     "soi"
```

Figure 14 : Liste des "Stopwords" de tm (en)

```
> stopwords("en")
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"
 [10] "your"       "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"    "she"
 [19] "her"        "hers"       "herself"    "it"         "its"        "itself"     "they"       "them"       "their"
 [28] "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"        "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"         "does"       "did"        "doing"      "would"
 [55] "should"     "could"      "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"
 [64] "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"    "isn't"
 [82] "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"
[100] "that's"     "who's"      "what's"     "here's"     "there's"    "when's"     "where's"    "why's"      "how's"
[109] "a"          "an"         "the"        "and"        "but"        "if"         "or"         "because"    "as"
[118] "until"      "while"      "of"         "at"         "by"         "for"        "with"       "about"      "against"
[127] "between"    "into"       "through"    "during"     "before"     "after"      "above"      "below"      "to"
[136] "from"       "up"         "down"       "in"         "out"        "on"         "off"        "over"       "under"
[145] "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"      "why"
[154] "how"        "all"        "any"        "both"       "each"       "few"        "more"       "most"       "other"
[163] "some"       "such"       "no"         "nor"        "not"        "only"       "own"        "same"       "so"
[172] "than"       "too"        "very"
```

[10] Pour afficher cette liste, soumettre dans la fenêtre R Commander la commande stopwords("fr") ou stopwords("en") selon la langue.