## Utilisation d'un outil de statistiques textuelles<sup>1</sup>

# IRaMuteQ 0.7 alpha 2 Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires

### Elodie Baril et Bénédicte Garnier

Institut National d'Etudes Démographiques Paris (France)

Logiciel libre développé par Pierre Ratinaud.



### http://www.iramuteq.org/

Les données utilisées dans ce support sont extraites du projet EuroBroadMap (<u>http://www.eurobroadmap.eu/</u>).

<sup>&</sup>lt;sup>1</sup> Ce document ne remplace pas un guide d'utilisation du logiciel mais donne des indications sur les menus qui nous ont semblé utiles pour analyser des données textuelles.

Nous remercions France Guérin-Pace de nous avoir fait partager ses premiers retours d'expérience sur l'utilisation d'IRaMuteQ.

Nous traitons les réponses des étudiants interrogés en Chine à une question ouverte posée comme suit : « Quels sont les mots que vous associez le plus à l'« Europe » ? Choisissez 5 mots au maximum ». Les réponses sont de l'ordre de quelques mots.

### Table des matières

1 / Importer le fichier à analyser	4
2 / Statistiques	8
3 / Spécificités et AFC	12
4 / Classification	19
5 / Analyse de similitudes	24
6 / Nuage de mots	27
7 / Création de sous corpus	29
Références	
Table des figures	

### Installer IRaMuTeQ

Le logiciel est gratuit, il faut le télécharger à partir du site http://www.iramuteq.org/telechargement.

Il nécessite également l'installation d'une version récente du logiciel R (et de préférence la version 3.1).

### Tutoriel

Une documentation sur le formatage des corpus texte est disponible sur le site du logiciel (<u>http://www.iramuteq.org/documentation/formatage-des-corpus-texte</u>).

### Les résultats

Les résultats des calculs des différentes étapes de l'analyse textuelle sont sauvegardés au fur et à mesure de l'exécution dans des sous-répertoires par type d'analyse. On y retrouve des fichiers (au format .csv) et des graphiques.

Les analyses sont sauvegardées dans un fichier (.ira).

Avant toute mise en œuvre, au vu du nombre de répertoires et de fichiers générés, il est recommandé de déposer le fichier correspondant au document à analyser dans un répertoire dédié.

## Description des menus d'IRaMuTeQ

🐕 IRaMuTeQ 0.7 alpha 2	
Fichier Edition Vue Analyses de	matrice Analyse de texte Aide
📳 🔳 🔒 😵 🗖 😐 🎯	
Historque Corpus textuel	IRaMuTeQ
	http://www.iramuteq.org
	Version 0.7 alpha 2
	Laboratoire LERASS
	REPERE Licence GNU GPL (c) 2008-2014 Pierre Ratinaud
	python <sup>TH</sup> Le propos de cet ouvrage est de fixer le peu graphique fie nor spece fund de fixer le peu graphique fie nor spece fund de Verre happe pas que cet essai va, ou du moins semb
Prêt	Bienvenue

## Edition → Préférence : Options du logiciel

Paramètres	TDoMar	×
Jouer un son à la fin des analyses		⊘ oui ⊚ non
Vérifier l'existence d'une nouvelle version au démarrage		◉ oui ⊘ non
Langue de l'interface		french 🔹
Vérifier l'installation des libraires de R		Vérifier
Utiliser svdlibc	Chemin : /usr/bin/svd	Browse
Chemin de R	Chemin : C:\Program Files\R\R-3.1.0\bin\x64\	Browse
Miroir par défaut de R	http://cran.rstudio.com/	
Mémoire maximum de R	×.	
	ОК	Cancel

## 1 / Importer le fichier à analyser

Dans ce document, nous ne traitons pas les menus *ouvrir une matrice* (importation d'un tableau de données comportant des valeurs numériques) ni *importer de* TXM (plateforme logicielle open-source pour la textométrie, voir <u>http://textometrie.ens-lyon.fr/</u>) ou *importer depuis Europress* (données d'information de presse <u>http://www.bpe.europresse.com/</u>) et *Factiva* (données issues de média comme des journaux, magazines retranscriptions radio et télévision, photos, etc..).

*Ouvrir une analyse* permet de récupérer des traitements (.ira) et d'accéder aux résultats calculés par le logiciel dans chaque sous dossier (onglet vue).

Le menu *Outils* permet de créer des sous corpus. Nous aborderons cette fonctionnalité à la fin du document.

#### Figure 1 - Menu Fichier



**Ouvrir un corpus texte** permet de charger un fichier texte qui respecte la mise en forme « Alceste »<sup>2</sup> comportant des lignes étoilées entre chaque réponse. Les textes à analyser sont très courts et on dispose de caractéristiques sur ces textes (comme le pays de naissance des enquêtés, la ville d'enquête, le sexe, etc.)(Figure 2)

#### Figure 2 - Extrait du fichier traité (EBM\_iram\_CHN\_n.txt)

```
0241 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc1
clean fashionable healthy civilized
0242 *p_CHN *v_BJS *s_F *d_ART *e_0 *r_Inc3
developed economy beauteous environment linguistic diversity
0244 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc2
small area small population good environment beautiful scenery
0245 *p_CHN *v_BJS *s_M *d_ART *e_2 *r_Inc3
gleichschaltung contradiction civilized bright future
0246 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc2
free developed democratic lodgeable
0247 *p_CHN *v_BJS *s_M *d_ART *e_2 *r_Inc3
small cold foolish leisure expensive
```

<sup>&</sup>lt;sup>2</sup> La première ligne introduit chaque texte à analyser (exemple : une réponse à une question ouverte) par les caractéristiques du locuteur. En premier lieu, figure l'identifiant du texte, suivi d'une série de modalités de variables qualitatives précédées d'une étoile et du nom de la variable. Il est préférable de mettre un blanc souligné entre le nom de la variable et la modalité pour qu'il soit possible par la suite d'extraire des sous-corpus selon les modalités d'une de ces variables (Garnier, Guérin-Pace, 2010). La ligne étoilée peut débuter par 4 chiffres (identifiant de questionnaire ou de texte par exemple) ou 4 étoiles.

Le texte à analyser ne doit pas comporter d'étoiles car le caractère \* est réservé aux caractéristiques sur les textes dans la *ligne étoilée* dédiée.

Une fois de nom du fichier renseigné, IRaMuteQ propose de paramétrer la transformation du corpus pour effectuer l'analyse de texte.

### La création du lexique

 $\rightarrow$  Ouverture de la fenêtre de paramétrage avant le lancement de l'analyse (Figure 3). Par défaut, IRaMuteQ fait appel à des dictionnaires de la langue française mais si le corpus à analyser est dans une autre langue, changer alors le paramètre *Langue* (ici le texte est en anglais). Cela sera important pour la reconnaissance des catégories de mots (dans la phase de lemmatisation). Il est possible aussi de changer le répertoire destination des résultats *Répertoire en sortie*.

Le *marqueur de texte* correspond au séparateur entre unités d'analyses (ici les 1140 réponses des étudiants interrogés en Chine). Nous utilisons ici les identifiants des questionnaires (codés sur 4 caractères).

Le *dictionnaire d'expression* repère des expressions courantes comme « aujourd'hui » ou « grand-père » (en français).

*Faire des segments de texte* permet à IRaMuteQ de découper les textes longs (cas d'entretiens par exemple) en unités plus petites (les segments de texte). Ce découpage peut se faire en fonction d'un nombre d'occurrences (cas par défaut), d'un nombre de caractères ou de paragraphes.

Bare o mackation aa torr	745
Paramètres	×
Général Nettoyage	
corpus	D:\ADTextuelles\Test_Iramuteq\Tests\EBM\EBM_iram_CHN_n.txt
Nom du corpus	EBM_iram_CHN_n_
Encodage	cp1252 - Western Europe 🔹
Langue	english
	Défaut     english
Dictionnaire	O Autre Browse
Répertoire en sortie	D:\ADTextuelles\Test_Iramuteq\Tests\EBM\EBM_iram_CHN_n_corpus_V7 Changer
Marqueur de texte	0000 -
Utiliser le dictionnaire des expressions	
Faire des segments de texte	
Méthode de construction des segments	occurrences 🔻
Taille des segments de texte	40

#### Figure 3 - Indexation du corpus

#### Avril 2015

Par défaut, IRaMuteQ transforme tout le texte en minuscules pour ne pas différencier les mots écrits tout en minuscules des mêmes mots écrits avec une majuscule en début de phrase (Figure 4).

ingure 4 - Options du « Nettoyage » automatique du nemer
Paramètres
Général Nettovage

Figure 4 - Options du « Nettoyage » automatique du fich	nier
---	------

Général Nettoyage	
Mettre le texte en minuscule	
Éliminer les caractères en dehors de cette liste	✓ ^a-zA-Z0-9àÀâÂäÄáÁåÅãéÉèÈêËïÌÌÌÌĨIÍIÓÓÒÒôÔöÖöÖöÖøØùÙûÛüÜúÚçÇBc
Remplacer les apostrophes par des espaces	
Remplacer les tirets par des espaces	
Conserver la ponctuation	
Pas d'espace entre deux formes	

Une fois le corpus indexé, IRaMuteQ affiche une première description quantitative du texte (Figure 5).

Ici on dénombre 1140 segments de texte, correspondant au nombre de textes initial car le corpus n'a pas été découpé et 1138 formes graphiques (ici des formes/mots) différents.

R a créé un premier tableau lexical croisant les textes et les formes (Document Term Matrix du package tm de R<sup>3</sup>).

х

<sup>&</sup>lt;sup>3</sup> http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

#### Figure 5 - Bilan lexical

Description EBM_iram_CHN_n_ ×		
Description du corpus		
Nom	EBM_iram_CHN_n_	
Langue	english	
Encodage	cp1252	
originalpath	D:\ADTextuelles\Test_Iramuteq\Tests\EBM\EBM_iram_CHN_n.txt	
pathout	D:\ADTextuelles\Test_Iramuteq\Tests\EBM\EBM_iram_CHN_n_corpus_V7	
date	Tue Mar 24 16:57:16 2015	
time	0h 0m 1s	
Paramètres		
ucemethod	1	
ucesize	40	
keep_caract	^a-zA-Z0-9àÀâÂäÄáÁâÁãéÉèÈêÊëËìÌÌÌÌĬĴÓÓòÒôÔöÖõÕõØøÙùÙûÜüÜúÚçÇßœŒ'ñÑ.:,;!?'	
expressions	1	
Statistiques		
Nombre de textes	1140	
Nombre de segments de texte	1140	
occurrences	5095	
Nombre de formes	1138	
Nombre d'hapax	674 - 59.23 % des formes - 13.23 % des occurrences	

### Analyse de Texte

IRaMuTeQ propose différents types d'analyses (Figure 6) basées sur : la lexicométrie (*Statistiques*), les méthodes statistiques (calcul de *Spécificités, analyse factorielle* ou *Classification*), la visualisation de données textuelles (*Nuage de mots*) ou l'analyse de réseaux de mots (*Analyses de similitudes*).

Figure 6	5 -	Menu	Analyse	de	texte
----------	-----	------	---------	----	-------

🔓 IRaMuTeQ 0.7 alpha 2				
Fichier Edition Vue Analyses de	matrice	Anal	yse de texte Aide	
i 📕   📕   🎭   😵   褌   📾 🎯	) 🔞 🛛	1	Statistiques	
Historique	Desc	-	Spécificités et AFC	
Recueil_test_corpus_1			Classification	•
Theatre_corpus_2	Descrip	XL	Analyses de similitudes	
	Nom	-	Analyses de similades	
⊕ Theatre corpus 1		-	Nuage de mots	
Paris_PDP_ALC_corpt	Langue		Sous-corpus	•
	Encoda	*	Exporter la table des méta-données	
	. · ·	_		_

## 2 / Statistiques

Dans cette partie, IRaMuTeQ affiche tout le lexique du corpus.

Il faut d'abord choisir de lemmatiser ou non les formes/mots et paramétrer les catégories de mots à prendre en compte dans les calculs (Figure 7).

Figure 7 -	Paramétrage de	la lemmatisation
------------	----------------	------------------

Paramètres		
Lemmatisation	◉ oui ⊘ non	
Paramètres des clés	propriétés	Équivalent des clés catégorielles d'Alceste
Dictionnaire	<ul><li>indexation</li><li>Autre</li></ul>	
Chemin		Browse
	ОК	Cancel

Le logiciel fait une *lemmatisation* (Figure 7) à l'aide de ses dictionnaires<sup>4</sup> et peut ainsi regrouper les formes au singulier et au pluriel sous une même forme, les verbes conjuguées sous la forme infinitive.

L'indexation à l'aide de dictionnaire permet aussi à IRaMuTeQ d'identifier les expressions et les catégories grammaticales des mots pour leur attribuer une clé d'analyse. Selon cette clé, il les traitera en élément *actif* ou *supplémentaire* (Garnier, Guérin-Pace, 2010) dans les analyses ou le découpage du corpus en segments de texte.

*Paramètres des clés*  $\rightarrow$  propriétés : permet de modifier les clés d'analyse par catégories et de différencier le traitement de certaines formes (Figure 8).

<sup>&</sup>lt;sup>4</sup> Dictionnaires anglais, allemands, italiens, espagnols, portugais (certains sont encore expérimentaux), dictionnaires minimalistes pour le suédois et le grec.

Clés d'analyse		-			×					
Choix des dés d'analyse 0=éliminé ; 1=active ; 2=supplémentaire										
Adjectif	þ	voir liste	Conjonction	2	voir liste					
Adjectif démonstratif	2	voir liste	Formes non reconnues	1	voir liste					
Adjectif indéfini	2	voir liste	Nom commun	1	voir liste					
Adjectif interrogatif	2	voir liste	Nom supplémentaire	2	voir liste					
Adjectif numérique	2	voir liste	Onomatopée	2	voir liste					
Adjectif possessif	2	voir liste	Pronom démonstratif	2	voir liste					
Adjectif supplémentaire	2	voir liste	Pronom indéfini	2	voir liste					
Adverbe	1	voir liste	Pronom personnel	2	voir liste					
Adverbe supplémentaire	2	voir liste	Pronom possessif	2	voir liste					
Article défini	2	voir liste	Pronom relatif	2	voir liste					
Article indéfini	2	voir liste	Préposition	2	voir liste					
Auxiliaire	2	voir liste	Verbe	1	voir liste					
Chiffre	2	voir liste	Verbe supplémentaire	2	voir liste					
				-	Vol 11502					
			Apply							

#### Figure 8 - Paramétrage de l'analyse des formes par catégories grammaticales

- Ce qui est mis en **actif** par défaut (codé 1): adjectifs, adverbes, formes non reconnues, noms communs et verbes.

- Ce qui est mis en supplémentaire par défaut (codé 2): mots outils.

Attention l'option « voir liste » affiche des exemples qui ne correspondent pas aux mots du corpus analysé.

Un mot qui n'est pas dans le dictionnaire est mis dans la catégorie *Formes non reconnues*. Il est possible de l'ajouter dans le dictionnaire et y indiquer sa catégorie grammaticale.

### Modifier le(s) dictionnaire(s)

- Aller dans le répertoire de l'environnement utilisateur
- Ex : C:\Users\garnier\.iramuteq\dictionnaires
- Copier le dictionnaire correspondant à la langue (ex : lexique\_fr.txt) et donner un nom différent à l'Initial (ex : lexique fr ini.txt)

Extrait du dictionnaire français

ôtes ôtez ôtiez ôtât ôtât ôtêrent ôté ôtée ôtées ôtés	ôter ôter ôter ôter ôter ôter ôter ôter	ver ver ver ver ver ver ver ver ver	m f f m	s s p p	$\begin{array}{c} 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ 16.81\\ \end{array}$	42.03 42.03 42.03 42.03 42.03 42.03 42.03 42.03 42.03 42.03 42.03	0.65 1.3 0.17 0.02 0 3.18 0.42 0.16 0.04	0 0.81 0 0.07 0.14 0.27 5.47 0.54 0.07 0.14	<pre>ind:pre:25; ind:pre:25; ind:inp:20; ind:pre:10; sub:inp:35; ind:pas:30; par:pas; par:pas; par:pas; par:pas; par:pas; par:pas;</pre>	•
· ·									,	af .

Ajouter une ligne pour chaque nouvelle forme et renseigner au moins les trois premières colonnes (1ère colonne : forme initiale, 2ème colonne : forme racine et 3ème colonne catégorie/clé d'analyse)

Par défaut, les termes non reconnus sont mis dans la catégorie Forme non reconnue (nr) et traités en actif si on laisse le paramétrage par défaut de la lemmatisation.

Si on veut qu'un mot nouveau soit traité en élément supplémentaire, il faut le mettre dans une catégorie traitée en supplémentaire (ex : Conjonction)

Une fois le paramétrage validé (OK), IRaMuTeQ affiche les résultats (Figure 9, Figure 10) et génère un répertoire (ou dossier) dans lequel il place des fichiers résultats : *nomdufichier texte\_*stat\_1.

Pour toutes les analyses, un clic droit sur une analyse permet d'afficher les options utilisées pour le traitement.

Il est également possible d'exporter le dictionnaire d'un corpus ou le dictionnaire des termes/mots à partir d'une analyse statistique (Figure 9).

Figure 9 - Bilan lexical de la première analyse du corpus (onglet résumé)



(EuroBroadMap 2009)

Figure 10 - Affichage du lexique des formes actives

Descripti	on EBM_iram_CHN	V_n_	EBM_irar	n_CHN_n_stat_2 >	EBM_iram_Cl			
Résumé	Formes actives	×	Formes supplémentaires Total Hapax					
Forme			Freq. 🖖	Types				
develop	ed		355	ver				
rich			244	adj				
romanti	ic		204	nom				
beautifu	l l		166	adj				
advance	ed		96	ver				
civilized			78	ver				
small			66	adj				
freedom	n		65	nom				
open			65	adj				
classica	I		64	nom				
high			62	adj				
good			59	adj				
elegant			57	nr				
civilizati	ion		51	nom				
environ	ment		47	nom				
clean			41	adj				
welfare			41	nom				
countrie	25		40	nom				
democr	acy		39	nom				
leicure			36	nom				

1<sup>er</sup> onglet
 Résumé = description générale du corpus (nombre de textes, d'occurrences, de formes...)

2<sup>ème</sup> onglet • Formes actives = liste des formes/mots actifs (avec leur catégorie grammaticale) par fréquences décroissantes.

3<sup>ème</sup> onglet • Formes supplémentaires = liste des formes/mots supplémentaires par fréquences décroissantes

4<sup>ème</sup> onglet • Total = ensemble des mots par fréquences décroissantes

5<sup>ème</sup> onglet • Hapax = mots du corpus présents une seule fois

Sur chaque forme/mot

Clic droit  $\rightarrow$  Formes associées permet de visualiser les regroupements (lemmatisation) Clic droit  $\rightarrow$  Concordancier affiche le contexte d'utilisation du mot dans le corpus

Fichiers générés de le dossier « nomducorpus\_Stat\_1 »:

- analyse.ira : Fichier permettant d'ouvrir l'analyse déjà faite dans le logiciel.

- formes\_actives (csv) : 3 colonnes avec une ligne par mots que le logiciel prend en compte ; leur fréquence, la catégorie du mot.

- formes\_supplémentaires (csv) : mots non pris en compte ; fréquence ; type :préposition (pre), adj\_pos, art\_def, adj\_pos art\_ind, conjonction (con), pro\_per, art\_ind, art\_def, aux (auxiliaire), num (chiffre), pro\_dem, pro\_ind, pro\_rel, ver\_sup (vouloir, devoir, faire, pouvoir...), ono (derrière, dehors, pouce).

- glob (txt) : fichier Global : nombre d'uci : ici 1140 ; nombre d'occurrences : 5095 ; nombre de formes : 1729 ; moyenne d'occurrences par forme : 4.65 ; nombre d'hapax : 634 (12.44% des occurrences - 33.69% des formes) ; moyenne d'occurrences par uci : 4.47

- hapax (csv) : mots ayant une fréquence de 1.

- total (csv) : Tous les mots, fréquences décroissante à partir de 2 occurrences.

 $\rightarrow$  Permet de visualiser les mots non lemmatisés et leur catégorie.

- Zipf : graphique présentant en ordonnée les fréquences et en abscisse les rangs des formes du corpus.

## 3 / Spécificités et AFC

Cette analyse permet d'identifier les mots spécifiques par sous-catégories et réalise une Analyse Factorielle sur un tableau lexical agrégé (TLA) construit avec les variables sélectionnées.

Choix des variables	×	Choix des variables	×
Formes utilisées	actives et supplémentaires 🔻	Formes utilisées	actives et supplémentaires 🔻
Sélection par	variables 🔻	Sélection par	modalités 🔻
Choix	*v *p *r *s *d *e	Choix	*d_BUS *d_ENG *d_HEA *d_POL *d_SHS *e_0 *e_1 *e_2 *e_3 *e_4 *e_5 *p_CHN *r
Indice	loi hypergéométrique 🔻	Indice	loi hypergéométrique 👻
Fréquence minimale	10 🛋	Fréquence minimale	10 🛋
	Cancel OK		Cancel OK

Figure 11 - Sélection par type de variables ou sélection par modalités

En sélectionnant par variables (Figure 11) on ne peut choisir qu'une variable à la fois (celle qui est sélectionnée en premier) et IRaMuTeQ ne fait pas d'AFC avec une variable qui a trop peu de modalités ; on peut ne sélectionner que les formes actives ou supplémentaires. En faisant une sélection par modalités, on peut choisir plusieurs variables d'intérêt et retirer les modalités rares (peu d'individus).

Choix des variables/modalités pour calculer les spécificités et construire le tableau lexical

EBM_iram_CHN_n_stat_1	Spécificité	is - El	BM_iram_CHN	_n_spe	:_1 ) s	Spécificités - EBM_iram_CHN_n_spec_2 ×			
Formes × Formes banales	Types	Fréq	juences des for	mes	Fréque	nces des types	Fréquences	relatives des	
formes	*v_BJS	∿	*v_CAN	*v_N	KG	*v_SHA	*v_WUH		
football	1.6901		-0.2606	-0.73	03	0.459	-1.1509		
small	1.5102		-0.3163	-2.26	62	0.3468	0.4345		
open	1.2917		-1.9916	0.537	7	-0.6227	0.6212		
good	1.0993		-0.3853	-0.82	57	0.7514	-0.7337		
and	0.9451		0.3746	-0.26	14	-1.439	0.4079		
a	0.9195		0.3988	-0.32	03	-0.5027	-0.4799		
democratic	0.8527		0.3596	-0.35	43	-1.2126	0.4711		
modernization	0.8424		0.6438	0.364	18	-0.7006	-1.0988		
quality	0.8065		0.3335	-0.40	18	-0.3678	-0.3626		
noble	0.7438		-0.3263	-0.29	08	0.2195	-0.2679		
european	0.7438		-0.6736	0.613	31	0.2195	-0.5821		
high	0.7332		0.3829	-0.71	73	0.6176	-0.8502		
flourish	0.7096		-1.0691	-0.55	21	-0.7416	1.9332		
graceful	0.6923		-0.5791	0.313	36	-0.2515	-0.259		
comfortable	0.6835		-0.5568	0.314	16	-1.0651	0.8381		
classical	0.6584		0.459	-1.34	26	0.2894	0.3547		
union	0.6582		0.2348	0.266	53	-0.2266	-0.6493		
united	0.6582		-1.4221	-0.68	94	-0.2266	1.464		
capitalism	0.6291		-0.2213	0.465	58	-0.771	-0.2847		
rich	0.5966		-0.941	-0.27	42	0.6229	-0.3452		
fashion	0.5789		-0.4158	0.646	51	0.2578	-0.8207		
population	0.5326		-0.2792	-0.23	14	0.3759	-0.565		
cultural	0.5258		1.1392	-0.66	58	-0.5969	-0.3626		

#### Figure 12 - Mots spécifiques d'étudiants interrogés dans différentes villes chinoises

(EuroBroadMap 2009)

Plus la valeur est élevée (en valeur absolue), plus la forme/mot est spécifique de la modalité. Le signe + signifie que le mot est plus cité par ce groupe (ici étudiants interrogés à Pékin) que par les autres, le signe – signifie que le mot est moins cité par ce groupe que par les autres.

1 <sup>er</sup> onglet	• Formes (mots) : Affichage des formes spécifiques par modalité 🆑 et par
spécificité dé	écroissante (ici formes spécifiques des étudiants interrogés à Pékin v_BJS)
2 <sup>ème</sup> onglet	<ul> <li>Formes banales : Affichage des formes par effectif décroissant</li> </ul>

- 3<sup>ème</sup> onglet Types (adjectif, pronom...) : catégories grammaticales
- 4<sup>ème</sup> onglet Effectifs par formes/mots
- 5<sup>ème</sup> onglet Effectifs par types de catégories grammaticales

- 6<sup>ème</sup> onglet Effectifs relatifs des formes/mots
- 7<sup>ème</sup> onglet Effectifs relatifs par type grammatical de mot

8<sup>ème</sup> onglet • AFC (analyse factorielle des correspondances) sur un tableau lexical agrégé (TLA)

Sur chaque mot : Clic droit  $\rightarrow$  formes associées permet de visualiser les regroupements (lemmatisation) Clic droit  $\rightarrow$  concordancier affiche le contexte d'utilisation du mot dans le corpus Clic droit  $\rightarrow$  graphique affiche un graphique représentant le sur/sous emploi du mot Clic droit  $\rightarrow$  segment de texte caractéristique affiche des parties de textes spécifiques

- AFC forme : génère un graphique avec tous les mots analysés (Figure 13) et un graphique avec les variables étoilées.

- AFC type : génère un graphique avec le type des mots et un graphique avec les variables étoilées.

Figure 13 - Plan factoriel issu de l'AFC sur le Tableau Lexical Agrégé (mots et variable spécifique sélectionnée pour le calcul des spécificités)



(EuroBroadMap 2009)

En cliquant sur ce symbole on peut paramétrer le graphique des plans factoriels

R .	থ 🗙
Type de graphique	2D 💌
Format de l'image	png 🔻
Représentation	coordonnées 🔻
Variables	actives 🔻
largeur 800	hauteur 800
Taille du texte	12
Prendre les x premiers points	50
Prendre les x premiers par classe	30
Limiter les points par le chi2 de liaison aux classes	4
Empêcher les recouvrements	
Taille du texte proportionnelle à la fréquence 🛛 🕅	min 5 ^ max 40 ^
Taille du texte proportionnelle au chi2	min 5 A max 40 V
Facteur x: 1 🚔 Facteur y:	2 Facteur z: 3
	100

#### Figure 14 - Paramétrage des options de graphiques issus d'AFC

Type de graphe : 2D ou 3D

Format de l'Image : Png (format image) ou Svg (format vectoriel)

Représentation : choix entre coordonnées et corrélation

Variables : choix des variables à représenter entre actives, supplémentaires, étoilées, classes Taille : variation de la taille des formes en fonction de sa fréquence ou du Chi2 Facteur : possibilité de choisir les axes factoriels à afficher (Facteurs 1-2 par défaut)

Remarque : il n'est pas possible de déplacer les mots du graphique pour une meilleure visibilité. Pour cela, il faut enregistrer le graphique au format vectoriel (svg) et le travailler avec un logiciel de dessin vectoriel (comme Inkscape<sup>5</sup> ou Illustrator).

Pour garder les mots qui ont les plus fortes contributions, relancer l'analyse à l'aide du symbole ci-dessus pour sélectionner « contributions » dans la représentation.

<sup>&</sup>lt;sup>5</sup> https://inkscape.org/fr/

On retrouve tous les calculs de formes spécifiques et de l'Analyse Factorielle des Correspondances (contributions, coordonnées, etc.) dans le répertoire généré par IRaMuTeQ.

Fichiers disponibles dans le répertoire (nomcorpus\_spec\_n) :

- **afcf\_col.csv** : Ligne : classes ; Colonnes : Coord. facteur ; Corr. facteur 1 à 6 ; COR -facteur 1 à 6 ; CTR facteur 1 à 6 (contribution) ; mass ; chi.distance ; inertie
- afcf\_col.png : image / graphique des modalités actives (var étoilées)
- **afcf\_facteur.csv** : Ligne : facteurs ; Colonnes : valeurs propres ; pourcentages ; pourcentage cumulés
- afcf\_row.csv : Ligne : les mots (ne garde que les mots de fréquence supérieure au seuil indiqué dans le paramétrage, 11 par défaut) ; Colonnes : Coord. facteur de chaque classe ; Corr. facteur jusqu'à 6 ; « COR -facteur 1 » jusqu'à 6 ; CTR -facteur 1 à 6 ; mass ; chi.distance ; inertie .
- afcf\_row.png : Graphique des mots
- **afct\_col.csv** : Ligne : classes ; Colonnes : Coord. facteur de chaque classe ; Corr. facteur 1 à 6 ; COR facteur à 6 ; CTR -facteur 1 à 6; mass ; chi.distance ; inertie
- afct\_col.png : Graphique des modalités actives
- afct\_facteur.csv : Ligne : facteur ; Colonnes : valeurs propres ; pourcentages ; pourcentage cumulés
- afct\_row.csv : Ligne : type de mots ; Colonnes : « Coord. facteur de 1 à 6 ; Corrélation facteur 1 à 6 ; contribution facteur 1 à 6 ; mass ; distance du chi2 ; inertie
- **afct\_row.png** : Graphique avec les types de mots
- Analyse.ira : analyse qui peut être ouvert avec le logiciel dans « ouvrir une analyse → ouvre les onglets résultats de « spécificité et AFC ».
- banalites.csv : lexique des formes/mots par effectif décroissant
- eff\_relatif\_forme.csv : Ligne : les mots ; Colonnes : les modalités des variables étoilées sélectionnées
- **eff\_relatif\_type.csv** : Ligne : les types de mots (24) ; Colonnes : les modalités des variables étoilées sélectionnées dans le paramétrage
- **tableafcm.csv :** Ligne : les mots retenus ; Colonnes : les modalités des variables étoilées sélectionnées (en effectif]
  - $\rightarrow$  Equivalent de l'onglet « Effectifs formes » dans le logiciel
- tablespecf.csv : Ligne : les mots ; Colonnes : les modalités des variables étoilées sélectionnées
   →Equivalent de l'onglet « formes » dans le logiciel : indique les termes les plus spécifiques de chaque modalité
- tablespect.csv : Ligne : les types de mots
   Colonnes : les modalités des variables étoilées sélectionnées
   tabletypem.csv : Ligne : les types de mots
   Colonnes : les modalités des variables étoilées sélectionnées (en effectif)

Il est possible de procéder à des regroupements de formes/mots appelés TGen à partir de la liste des formes/lemmes du corpus ou de l'onglet spécificités.

Figure 15 - Accès à l'éditeur de TGen dans la fenêtre historique d'IRaMuTeQ



Figure 16 - Accès au menu Faire un TGen à partir du menu Spécificités et AFC

alyses de matrice Analyse de texte Aide									
🔤 🎯	6	)   🧭   🏠 🗄	= : `	S 🗧	降 🏶 S	<b>S *</b> :	00		
		Description EBI	M_iram	_CHN_n_	_corpus_2 YE	BM_iram_C	CHN_n_st	tat_1 Cla	ssificatio
	•	Fréquences d	es fori	mes ×	Fréquences d	es types	Fréque	nces relativ	es des for
HN_n	fo	rmes 👍			*v_BJS	*v_(	C	*v_NKG	*v_SHA
HN_n		civilize			15	16	23		13
HN_n		classical			14	15	7		15
l_n_		clean			4	8	11		13
5		climate			6	11	4		4
us_2 ≡		cold			3	8	1		1
		comfortable			7	7 4 6			3
2		country			9	9		8	7
		cultural			6	9	3		4
AT_co		culture			3	5	6		3
AT_co		democracy			1	10	7		11
l_n_cc		democratic			Q	7	5		3
l_n_cc		develop		Formes a	ssociées				84
9		economic		Concord	ancier				3
8		economically		Graphiqu	e				3
5		economy		Segment	s de texte carac	téristiques	+		2
4		elegance	C	Faire un 1	IGen				2
lram_		elegant		. and an					13

lom Abund			
Formes			Sélection
formes 👍	eff	Types	r formes eff 👎
a	29	SW	abundance
abundance	1	nom	abundant
abundant	7	nr	abustle
abustle	1	nr	
academic	4	nom	
accidental	1	nom	
acquaintance	1	nom	
acracholia	1	nr	Add ->
active	2	nom	
admire	1	ver	<- Remove
advance	97	nom	
aegean	1	nr	
aesthetic	1	nom	
aestheticism	1	nom	
after	1	SW	
against	2	SW	
age	11	ver	
aggressive	2	nr	

Figure 17 - Création de types généralisés (TGen)

Ici on a choisi de regrouper abundance, abundant et abustle sous le TGen Abund

Il sera ensuite possible de visualiser les spécificités des types généralisés par sous-corpus et d'afficher le concordancier correspondant (Figure 18) (après avoir lancé le calcul).

#### Figure 18 - TGen spécifiques

Description EBM_iram_CHN	l_n_corpus_2 YE	BM_iran	n_CHN_n_stat_1	Classification -	EBM_iram_CHN	_n_corpus_2 Spé	cificités	s - EBM_iram_CHN_n_spec_	1
Fréquences des formes	Fréquences des ty	/pes Fréquences relatives des formes			Fréquences relatives des types AFC			Spécificités des TGens	×
formes	*v_BJS 👆	*v_CA	N *v_NKG	*v_SHA	*v_WUH				
Democrat	-0.8072	0.5701	-0.374	-0.3637	0.7734				
Abund	-0.315	0.2265	-0.336	-0.422	1.1099	_			
Abund       -0.315       0.2265       -0.336       -0.422       1.1099         Concordancier - Democrat       6335 *p_CHN *v_SHA *s_F *d_POL *e_0 *r_Inc3         mild and moist rich noble civilized ancient and democratic         1803 *p_CHN *v_CAN *s_M *d_BUS *e_0 *r_Inc1         democracy rich excellent surroundings         0246 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc2         free developed democratic lodgeable									

(EuroBroadMap 2009)

## 4 / Classification

#### Méthode Reinert

Implémentation de la méthode de classification « Alceste » de Max Reinert<sup>6</sup> (Figure 19).



Paramètres				
Classification	0	double sur simple sur s	RST segment	s de texte
	۲	simple sur t	texte	Nombre de formes pleines
Taille de rst1		12	·	tableau 1
Taille de rst2		14		Nombre de formes pleines
Nombre de classes terminales de la phase 1		10	<u>s</u>	classification double)
Nombre minimum de segments de texte par classe (0	= automatique)	0		Augmenter cette valeur
Fréquence minimum d'une forme analysée (2 = autor	natique)	2	*	modifie le nombre de classes
Nombres maximum de formes analysées		3000	*	
méthode pour svd	irlb	ba 🔻		
Mode patate (moins précis, plus rapide)				
Cancel Valeurs par	défaut OK	:]		

Remarque : on ne peut pas changer la « fréquence minimum d'une forme analysée » qui est en grisé. Seule la valeur du « nombre maximum de formes analysées » est prise en compte. Si le nombre total de formes actives est inférieur à cette valeur, seules les formes ayant un effectif d'au moins trois sont retenues.

Dans notre exemple, on a choisi d'opérer une classification simple sur textes car les textes (mots associés à « Europe ») sont très courts. Par défaut, la méthode propose de découper les textes en segments de textes en fonction du nombre de formes actives.

Pour afficher (et imprimer) le rapport d'analyse (équivalent du contenu de l'onglet « Profil ») faire un clic droit sur le nom de l'analyse correspondante (*nomcorpus\_*alceste\_n) de la fenêtre « Navigateur ».

<sup>&</sup>lt;sup>6</sup> Classification Descendante Hiérarchique de segments de textes à partir du Tableau Lexical Entier (Reinert, 1983).

Sortie résultats de la classification :

On trouve un résumé des résultats (nombre de textes, de formes, de classes, le pourcentage de textes classés et le dendrogramme) (Figure 20).

Figure 20 - 1er onglet : CHD

CHD × Profils AFC	
+-+-+-+-++   R a M u T e Q  - Thu Mar 26 12:47:13 2015 +-+-++-+-+-+++ Nombre de textes: 1140 Nombre de segments de texte: 1140 Nombre de formes: 1138 Nombre de formes: 1138 Nombre de lemmes: 1095 Nombre de formes actives: 1021 Nombre de formes actives: 1021	
Nombre de formes actives avec une fréquence >= 3: 276 Moyenne de formes par segment: 4.469298 Nombre de classes: 5	
917 textes classés sur 1140 (80.44%) ###################################	
classe 1	27.5 %
classe 5	

66	
	and the second s
	Party I

permet d'afficher les mots spécifiques des classes pour Cliquer sur les boutons aider à leur interprétation.

Ce dendrogramme peut être présenté pour montrer la répartition des classes les unes en fonction des autres.

IRaMuTeQ fournit pour chaque classe des aides à l'interprétation qui permettent à l'utilisateur d'appréhender l'univers lexical de la classe et de lui donner un intitulé/thème (Figure 21 - 2e onglet : Profils).

Pour chaque classe, on trouve les formes/mots les plus associés (effectifs, pourcentage, Chi2 d'association)

<ul> <li>Spécifici</li> </ul>	ités - EBM_iram_(	CHN_n_spec_2	Spécificit	és - EBM_i	ram_C	HN_n_spec	2_3 Classification - EBM	l_iram_CHN_n_ C
CHD Profils × AFC								
1 Classe 1 252/917 27.48%	2 Classe 2 270/917 29.44%	3 Classe 3 64/917 6.98%	4 Classe 4 250/917 27.26%	5 Classe 5 81/917 8.83%				Ŧ
num	eff. s.t.	eff. total	pourcentag	ge chi2	∿	Туре	forme	p 🔺
0	53	59	89.8	83 1	23.0	adj	high	< 0,0001
1	50	57	87.7	72 11	0.66	adj	small	< 0,0001
2	46	50	92	.0 11	0.46	adj	good	< 0,0001
3	37	39	94.8	87 9	2.82	nom	welfare	< 0,0001
4	41	46	89.1	13 9	92.36	nom	environment	< 0,0001
5	27	28	96.4	43 6	58 89	nom	country	< 0,0001
6	26	28	92	Form	les ass	ociées		< 0,0001
7	24	27	88	Chi2	par cla	isse		< 0,0001
8	23	26	88	Chi2	par cla	sse et den	drogramme	< 0,0001
9	15	15	10	Chi2	modal	lités de la v	ariable	< 0,0001
10	15	16	93	Grap	he du i	mot		< 0,0001
11	13	13	10	Con	Concordancier		< 0,0001	
67	30	49	61	Conc	oruan.	cier	,	< 0,0001
12	15	19	78	Faire	un IG	en		< 0,0001
13	12	14	85	Outil	s du C	NTRL (fran	çais uniquement) 🔹 🕨	< 0,0001
15	9	9	10	Gran	he de l	a classe		< 0,0001
14	9	9	10	Com	ine de l	4 - 444		< 0,0001
68	13	16	81	segn	ients r	epetes		< 0,0001
69	14	18	77	Segn	nents o	le texte car	actéristiques	< 0,0001
16	10	11	90	Nuag	je de n	nots de la o	classe	< 0,0001 🛫
Bienv	enue			Ехро	rter			
	27 50			Expo	rter po	ur Tropes		-
16	40 47	Chi2 par cla	sse + dendro	Ехро	rter po	ur Owledg	e	

#### Figure 21 - 2e onglet : Profils

 $\rightarrow$  Par clic droit sur les lignes (forme) on accède à d'autres menus offrant des aides à l'interprétation des classes (Figure 21):

- Formes associées au mot (si lemmatisation),

- Concordancier (dans les segments de texte de la classe, dans les segments de texte classés, dans tous les segments de texte),

- Outils du CNRTL : renvoie sur le site du Centre National de Ressources Textuelles et Lexicales et pour cette forme affiche (définition, étymologie, synonyme) si la langue du corpus est le français),

- segments répétés,

- segments de texte caractéristiques des classes. Choix entre 2 modes de calcul : *Absolu* (Somme des Chi2 des formes « marquées » du segment) (Figure 22) ou *Relatif* (moyenne des chi2 des formes marquées par segment).

Figure 22 - 3 réponses caractéristiques de la classe 2 (indice de rang absolu)



(EuroBroadMap 2009)

- Graphe du mot, graphe de la classe<sup>7</sup> (voir la partie Analyse de similitudes)
- Nuage de mot de la classe (voir la partie Nuage de mots)

- Exporter : génère le corpus correspondant à la classe (classe\_n\_export.txt) ou pour d'autres outils (comme Tropes ou Owledge)

IRaMuTeQ génère une analyse factorielle des correspondances (Onglet AFC) sur un tableau croisant la variable de classe et les mots/formes du lexique associé au corpus.





<sup>&</sup>lt;sup>7</sup> Attention : il est conseillé de faire les graphes dans l'ordre des classes: classe 1 puis classe 2 car dans les sorties, le fichier se nomme \_1, \_2 qui correspond à l'ordre de la création des graphes et non le numéro de la classe.

<u>Onglet AFC</u>: Le premier graphique correspond au plan factoriel (1-2) représentant les formes/mots actifs associés à « Europe » qui sont affichés de différentes couleurs selon la classe à laquelle elles appartiennent (5 classes dans notre exemple) (Figure 23). On retrouve aussi le pourcentage d'information résumée par chaque facteur (ici 31,65% pour le premier axe).

Le deuxième graphique représente les mots outils projetés en éléments supplémentaires sur ce même plan factoriel.

Le troisième représente les variables étoilées i.e les caractéristiques sur les textes introduites lors de la mise en forme du corpus. Dans notre exemple les différentes villes dans lesquelles étaient interrogés les étudiants (\*v\_BJS, \*v\_CAN, ...), leur domaine d'étude (\*d\_SHS, \*d\_HEA, ...). Ici aussi ces modalités sont projetées sur le plan factoriel de l'AFC croisant les formes actives et les modalités de la variable de classe.

Le quatrième graphique présente ici encore la projection des modalités de la variable classe (5 dans notre exemple).

<u>Onglet Facteurs</u> : On trouve les valeurs propres, pourcentage et pourcentage cumulées issus du calcul de l'AFC ;

Ici les fichiers générés sont sauvegardés dans un répertoire : nomcorpus\_alceste\_1 :

**Autres fichiers qu'on peut exporter** (par clic droit sur le nom du répertoire dans la fenêtre historique):

- profils des segments répétés (identique à l'onglet Profil de la CDH mais affiche les valeurs pour les segments répétés et non pour les mots)

- profils des types grammaticaux (pour repérer la sur-représentation de catégories grammaticales de formes par classes);

- exporter le corpus permet de générer un fichier html où chaque uce est associé à une couleur qui donne sa classe d'appartenance (corpus en couleur, Figure 24) et ainsi de repérer à quel numéro de classe correspond le segment de texte classé.

Figure 24 - Extrait du « corpus en couleur » issu d'une classification



## 5 / Analyse de similitudes

Il s'agit d'une analyse des cooccurrences présentée sous formes de graphiques de mots associés (Figure 25). Les indices de similitudes proposés dans IRaMuTeQ sont ceux disponibles dans la librairie *proxy* de R (Meyer, Buchta). Cette analyse met en œuvre les concepts de la théorie des graphes utilisés en Analyse des Données Relationnelles (Analyse de réseaux).





Par défaut, les formes/mots sont les sommets du graphe et les arêtes représentent les cooccurrences entre eux. Plus la taille des mots est grande, plus ils sont fréquents dans le corpus, plus les liens/arêtes sont épais, plus les mots sont cooccurrents.

Les options de taille et des couleurs des formes et les liens sont paramétrables dans l'onglet Paramètres graphiques.

Paramètres	-			Contraction of the local distance of the loc		
274			Paramètres du graphe	aramètres graphiques		
		Indice	cooccurrence 🔻			
formes	eff 🖖		Précentation	funchtannan arinanda 🚽		
developed	355	=	Fresentation			
rich	244		Type de graphique	Estationer - Format de l'image (une -		
romantic	204		Type de graphique	statique • Pormat de l'image prig •		
beautiful	166		Arbre maximum			
advanced	96					
civilized	78		Seuil pour les arêtes	1		
small	66					
freedom	65		Texte sur les sommets			
open	65		Indices sur les arêtes			
classical	64					
high	62		Arêtes courbées			
good	59					
elegant	57		Taille du texte	10		
civilization	51					
environment	47		Communautés 📃	edge.betweenness.community 💌 🔲 halo		
clean	41					
welfare	41		Sélectionnez une variable			

### Figure 26 - Paramétrage de l'analyse de similitude dans IRaMuTeQ Onglet Paramètres du graphe

Pour une meilleure visibilité, il est possible de sélectionner les mots selon leur fréquence (Figure 26, partie gauche de la fenêtre). Dans l'exemple ci-dessous, les mots ayant une fréquence supérieure à 6 ont été représentés.

L'algorithme de fruchterman reingold est utilisé pour optimiser l'affichage du graphe et visualiser les mots le plus « centraux » (mots « types » du corpus).

Les formes/mot mots les plus centraux sont détectés à partir du calcul de leur centralité d'intermédiarité. Ces mots servent d'intermédiaires pour relier (au sens de la cooccurrence) un grand nombre d'autres mots entre eux.

Il est possible d'exporter les graphes au format vectoriel (svg) ou pour gephi (format graphml) avec les coordonnées des points, la taille des sommets et leur couleur. (http://gephi.org)

Cocher « sélectionner une variable » permet de repérer les mots spécifiques de chaque modalité d'une variable. Par exemple pour la variable domaine d'études, les mots d'une même couleur (bleu clair, Figure 27) sont spécifiques de la modalité (d\_SHS).



Figure 27 - Extrait du graphe des mots associés à Europe par les étudiants interrogés en Chine selon le domaine d'études

La détection de communautés (Figure 26, onglet paramètres du graphe) est aussi une méthode issue de la théorie des graphes permettant de repérer des groupes (ici de mots). Les termes fortement coocuurents se trouveront au sein d'une même communauté et seront faiblement reliés avec les autres communautés (Figure 28).



Figure 28 - Détection de communautés dans le graphe des mots associés à "Europe"

### 6 / Nuage de mots

#### Figure 29 - Paramétrage du nuage de mots dans IRaMuTeQ

Préférences de wordcloud	23			
hauteur 800	largeur 800			
Format de l'image	png 🔻			
Nombre maximum de formes	600 ×			
Formes utilisées	actives			
Taille du texte	Min 5			
	Max 50			
Couleur du texte	Couleur du fond			
	OK Cancel			

On peut choisir de lemmatiser (ou non) le corpus, d'afficher les formes actives, supplémentaires ou les deux (Figure 29).

Cette analyse permet d'afficher le lexique des mots associés au corpus sur la forme d'un graphique appelé *Nuage de mots* où la taille des formes/mots est proportionnelle à leur fréquence. Les mots les plus cités sont placés au centre

#### Figure 30 - Nuage de mots associés à Europe par les étudiants interrogés en Chine



Pour plus de lisibilité, il est possible de sélectionner les mots à représenter (Figure 31).

rmes	eff 🦑	
modernization	12	
sea	11	
elegance	11	
powerful	11	
harmonious	11	
economic	10	
moderate	10	
world	10	
history	10	
west	9	
surroundings	9	
aging	9	
social	9	
easy	8	
human	8	
faddish	8	
gentle	8	
large	8	
atmosphere	8	
luxury	8	
arrogant	8	
·	0	

Figure 31 - Sélection de mots à afficher (critère de fréquence dans le corpus)

Figure 32 - Nuage des mots cités au moins dix fois



## 7 / Création de sous corpus

Les résultats issus de l'analyse de l'ensemble du corpus peuvent mettre en évidence la nécessité d'affiner l'exploration des données et de procéder à d'autres analyses sur des corpus plus restreints.

IraMuTeQ propose deux façons d'extraire des sous-corpus. Dans notre exemple, nous utilisons le menu *Sous-corpus par méta-données* car il permet d'utiliser les caractéristiques sur les textes introduits lors de leur mise en forme (variables étoilées) (Figure 33).

Sous-corpus	
Nom	SubEBM_iram_CHN_n_corpus_2
Sélectionnez une ou plusieurs méta-données	*d_ART *d_BUS *d_ENG # *d_ENG *d_HEA # *d_POL # *d_SHS *e_0 *-1
	*e_1 *e_2 *e_3 *e_4 *e_5 *p_CHN ▼
	OK Cancel

Figure 33 - Paramétrage de la sélection d'un sous-corpus par méta-données

Ici on ne garde que les réponses des étudiants en Sciences Humaines et Sociales (\*d\_SHS)

	0 1
Description du corpus	
Nom	SubEBM_iram_CHN_n_corpus_2
Langue	non défini
Encodage	cp1252
originalpath	D:\ADTextuelles\Test_Iramuteq\Tests\EBM\EBM_iram_CHN_n.txt
pathout	$\label{eq:linear} D: \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ $
date	Fri Apr 3 10:54:26 2015
time	0h 0m 0s
Paramètres	
ucemethod	non défini
ucesize	non défini
keep_caract	non défini
expressions	non défini
Statistiques	
Nombre de textes	668
Nombre de segments de texte	668
occurrences	2962
Nombre de formes	761
Nombre d'hapax	456 - 59.92 % des formes - 15.40 % des occurrences

Figure 34 - Bilan lexical du sous-corpus créé

Dans la partie « Classification », nous avons vu qu'il est également possible de créer un souscorpus contenant les segments de texte d'une classe.

Nous pouvons alors refaire les mêmes analyses disponibles dans le menu Analyse de textes sur ces sous-corpus.

## Références

- <u>http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation\_mashs2009.pdf</u>
- Reinert M. 1983, Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. Cahiers de l'Analyse des Données, 3, pp. 187-198
- <a href="http://www.eurobroadmap.eu/">http://www.eurobroadmap.eu/</a>
- Brennetot A., Emsellem K., Guérin-Pace F et Garnier B, « Dire l'Europe à travers le monde », Cybergeo : European Journal of Geography [<u>http://cybergeo.revues.org/25684</u>]
- Garnier B., Guérin-Pace F. 2010. Appliquer les méthodes de la statistique textuelle. Paris, CEPED, 86 p. (Les Clefs pour) [http://www.ceped.org/?Appliquer-les-methodes-de-la]

# Table des figures

Figure 1 - Menu Fichier	4
Figure 2 - Extrait du fichier traité (EBM_iram_CHN_n.txt)	4
Figure 3 - Indexation du corpus	5
Figure 4 - Options du « Nettoyage » automatique du fichier	6
Figure 5 - Bilan lexical	7
Figure 6 - Menu Analyse de texte	7
Figure 7 - Paramétrage de la lemmatisation	8
Figure 8 - Paramétrage de l'analyse des formes par catégories grammaticales	9
Figure 9 - Bilan lexical de la première analyse du corpus (onglet résumé)	10
Figure 10 - Affichage du lexique des formes actives	10
Figure 11 - Sélection par type de variables ou sélection par modalités	12
Figure 12 - Mots spécifiques d'étudiants interrogés dans différentes villes chinoises	13
Figure 13 - Plan factoriel issu de l'AFC sur le Tableau Lexical Agrégé (mots et variable	
spécifique sélectionnée pour le calcul des spécificités)	14
Figure 14 - Paramétrage des options de graphiques issus d'AFC	15
Figure 15 - Accès à l'éditeur de TGen dans la fenêtre historique d'IRaMuTeQ	17
Figure 16 - Accès au menu Faire un TGen à partir du menu Spécificités et AFC	17
Figure 17 - Création de types généralisés (TGen)	18
Figure 18 - TGen spécifiques	18
Figure 19 - Paramétrage de la méthode Reinert dans IRaMuTeQ	19
Figure 20 - 1er onglet : CHD	20
Figure 21 - 2e onglet : Profils	21
Figure 22 - 3 réponses caractéristiques de la classe 2 (indice de rang absolu)	22
Figure 23 - 3e onglet : plan factoriel (1-2) issu d'AFC représentant les formes actives	22
Figure 24 - Extrait du « corpus en couleur » issu d'une classification	23
Figure 25 - Extrait du graphe des mots associés à Europe par les étudiants interrogés en	
Chine	24
Figure 26 - Paramétrage de l'analyse de similitude dans IRaMuTeQ	25
Figure 27 - Extrait du graphe des mots associés à Europe par les étudiants interrogés en	
Chine selon le domaine d'études	26
Figure 28 - Détection de communautés dans le graphe des mots associés à "Europe"	26
Figure 29 - Paramétrage du nuage de mots dans IRaMuTeQ	27
Figure 30 - Nuage de mots associés à Europe par les étudiants interrogés en Chine	27
Figure 31 - Sélection de mots à afficher (critère de fréquence dans le corpus)	28
Figure 32 - Nuage des mots cités au moins dix fois	28
Figure 33 - Paramétrage de la sélection d'un sous-corpus par méta-données	29
Figure 34 - Bilan lexical du sous-corpus créé	29