

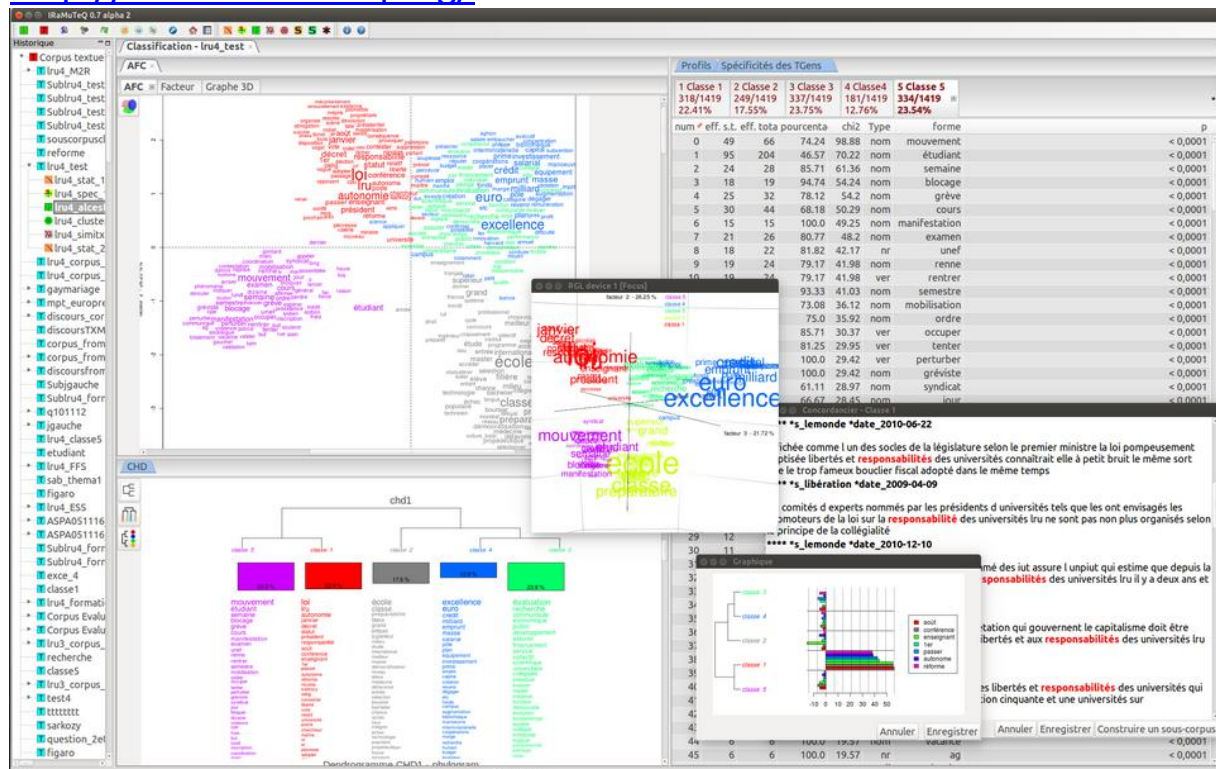
Utilisation d'un outil de statistiques textuelles¹

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 (<http://www.eurobroadmap.eu/>).

¹ 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.....	30
Table des figures.....	31

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 (<http://www.iramuteq.org/documentation/formatage-des-corpus-texte>).

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.

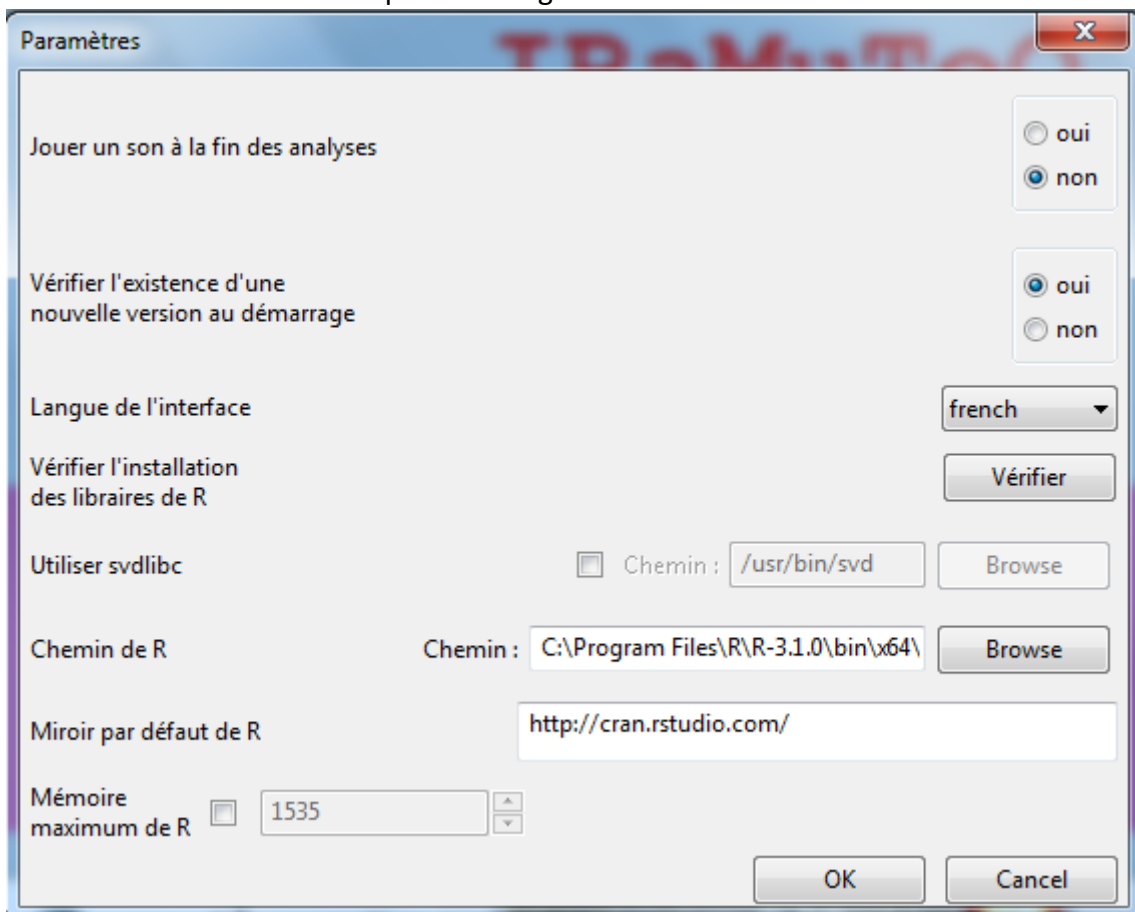
Les analyses sont sauvegardées dans un fichier (.ira).

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



Edition → Préférence : Options du logiciel



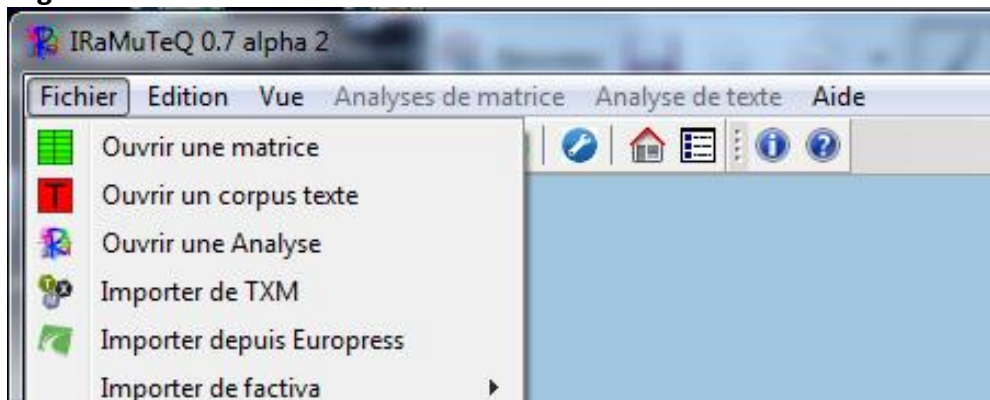
1 / Importer le fichier à analyser

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 <http://textometrie.ens-lyon.fr/>) ou *importer depuis Europress* (données d'information de presse <http://www.bpe.europresse.com/>) 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 »² 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
```

² 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 débiter 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

→ Ouverture de la fenêtre de paramétrage avant le lancement de l'analyse (Figure 3).

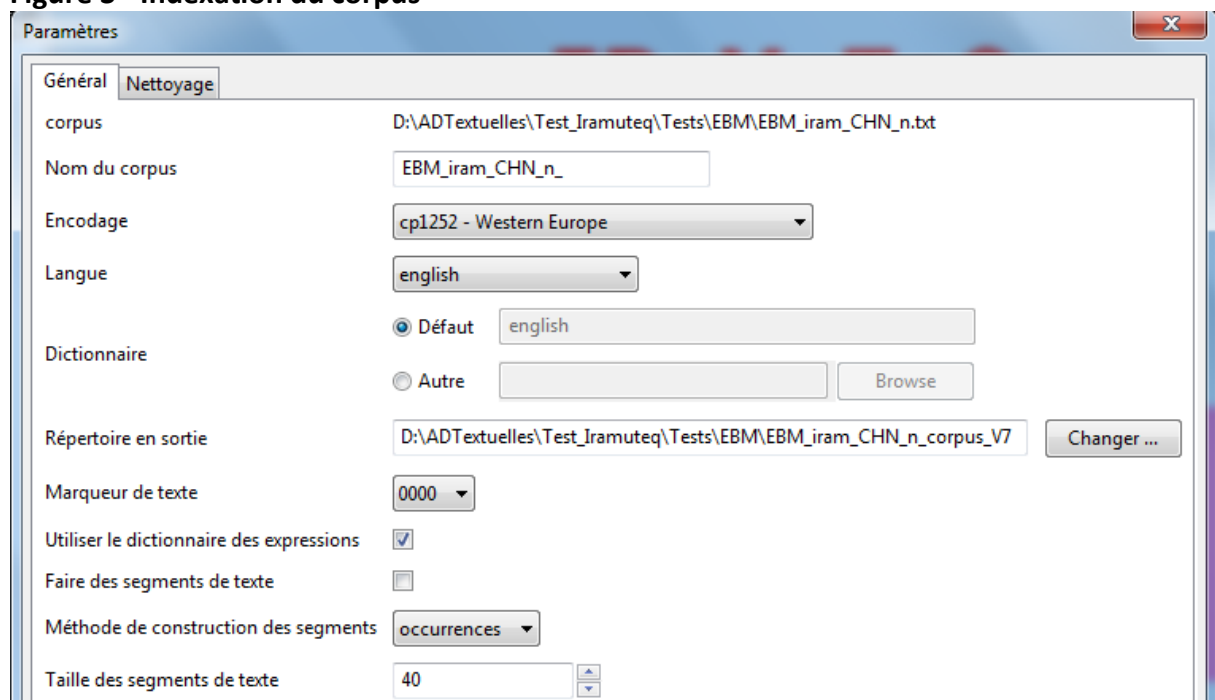
Par défaut, IRaMuteQ fait appel à des dictionnaires de la langue française mais si le corpus à analyser est dans une autre langue, changer alors le paramètre **Langue** (ici le texte est en anglais). Cela sera important pour la reconnaissance des catégories de mots (dans la phase de lemmatisation). Il 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).

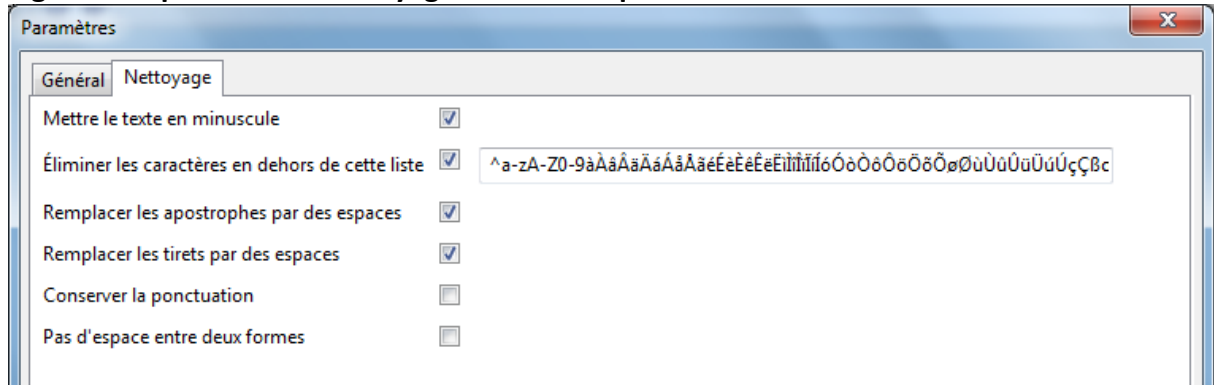
Faire des segments de texte permet à IRaMuteQ de découper les textes longs (cas d'entretiens par exemple) en unités plus petites (les segments de texte). Ce découpage peut se faire en fonction d'un nombre d'occurrences (cas par défaut), d'un nombre de caractères ou de paragraphes.

Figure 3 - Indexation du corpus



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).

Figure 4 - Options du « Nettoyage » automatique du fichier



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 R³).

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

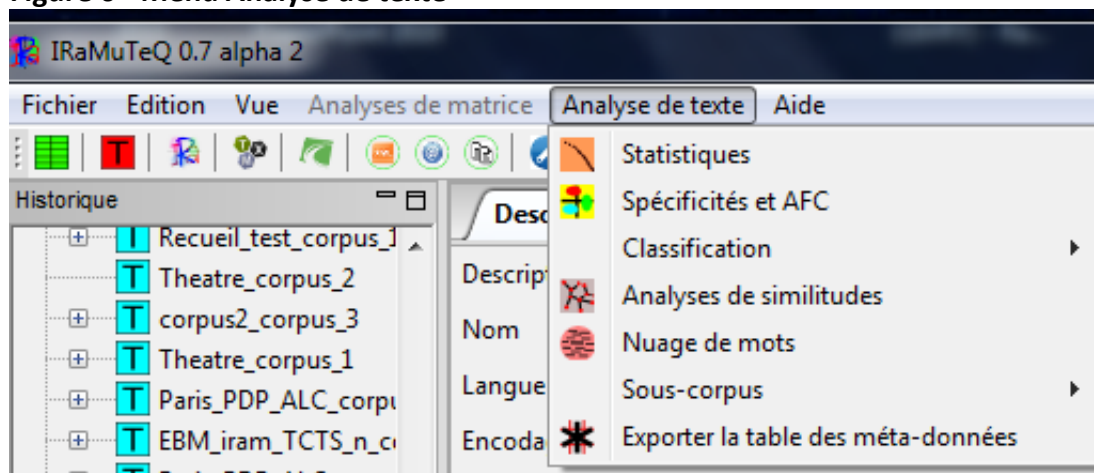
Figure 5 - Bilan lexical

Description EBM_iram_CHN_n_ x	
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-Z0-9àÁâÃäÅæÈéÊëËìíîïóÔõÖøÙúÛüÝçßœƒ'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 - Menu Analyse de texte

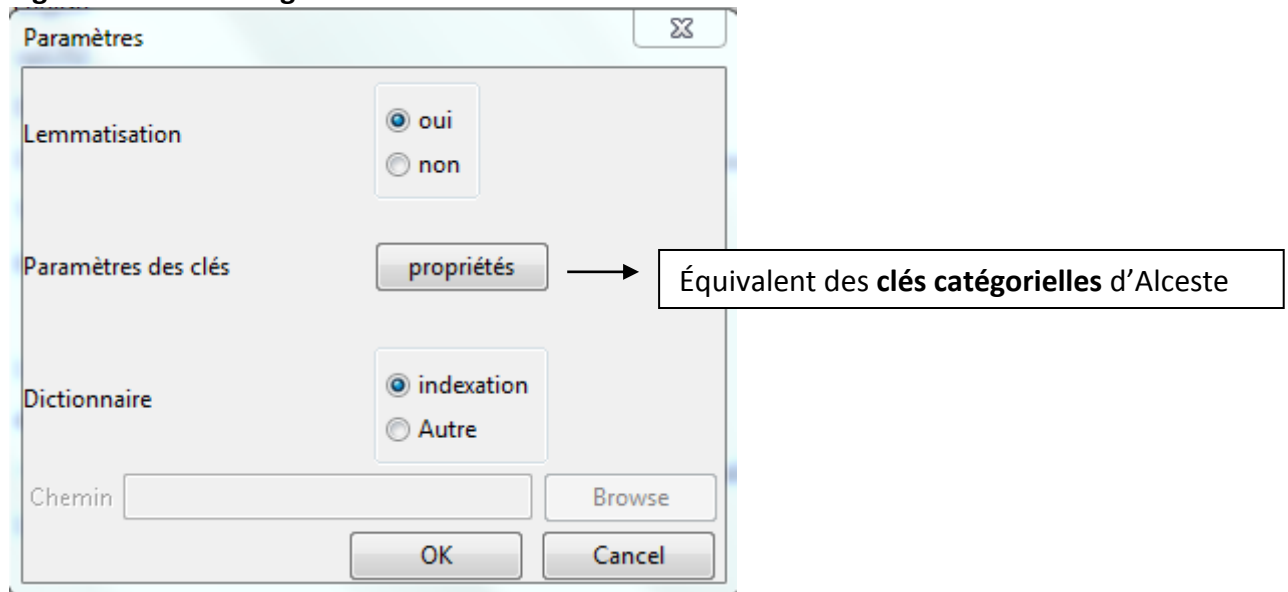


2 / Statistiques

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



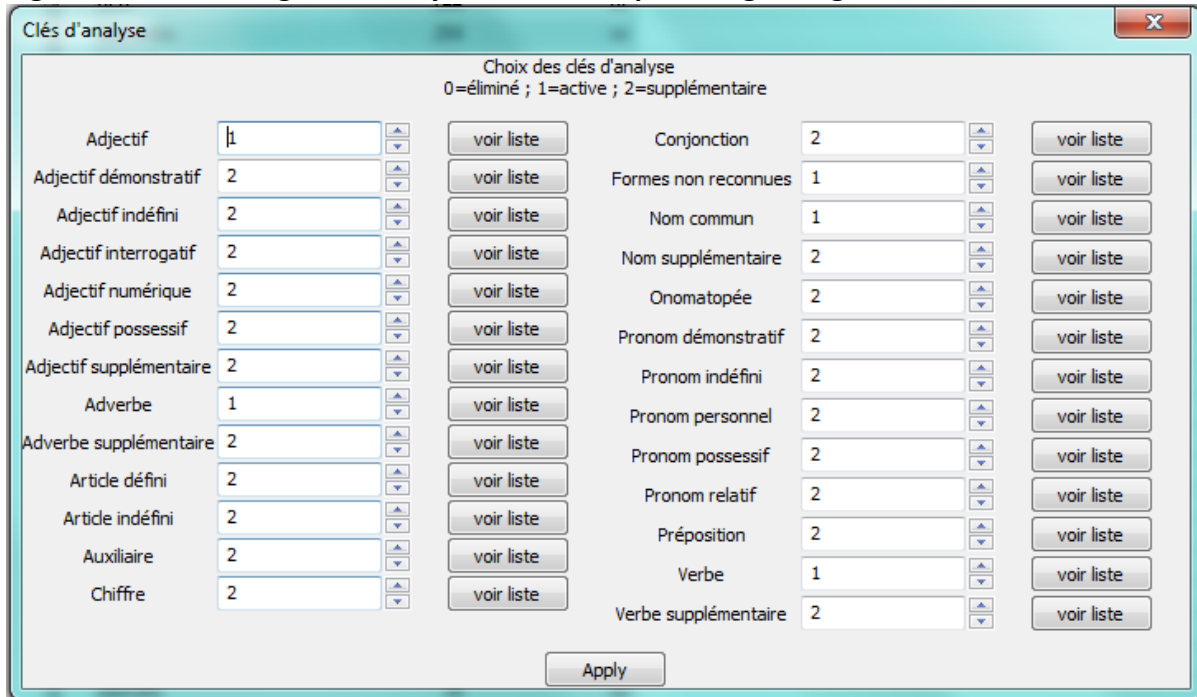
Le logiciel fait une *lemmatisation* (Figure 7) à l'aide de ses dictionnaires⁴ et peut ainsi regrouper les formes au singulier et au pluriel sous une même forme, les verbes conjugués sous la forme infinitive.

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 → propriétés : permet de modifier les clés d'analyse par catégories et de différencier le traitement de certaines formes (Figure 8).

⁴ Dictionnaires anglais, allemands, italiens, espagnols, portugais (certains sont encore expérimentaux), dictionnaires minimalistes pour le suédois et le grec.

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 ôter ver 16.81 42.03 0.65 0 ind:pre:2s;
ôtez ôter ver 16.81 42.03 1.3 0.81 imp:pre:2p; ind:pre:2p;
ôtiez ôter ver 16.81 42.03 0.17 0 ind:imp:2p;
ôtions ôter ver 16.81 42.03 0.02 0.07 ind:pre:1p;
ôtât ôter ver 16.81 42.03 0 0.14 sub:imp:3s;
ôtèrent ôter ver 16.81 42.03 0 0.27 ind:pas:3p;
ôté ôter ver m s 16.81 42.03 3.18 5.47 par:pas;
ôtée ôter ver f s 16.81 42.03 0.42 0.54 par:pas;
ôtées ôter ver f p 16.81 42.03 0.16 0.07 par:pas;
ôtés ôter ver m p 16.81 42.03 0.04 0.14 par:pas;
    
```

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)

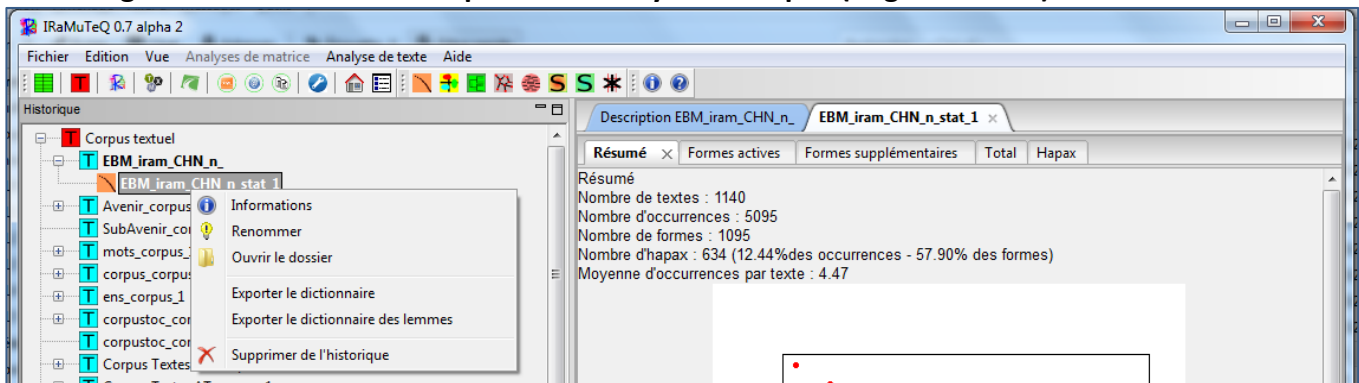
Par défaut, les termes non reconnus sont mis dans la catégorie *Forme non reconnue* (nr) et traités en actif si on laisse le paramétrage par défaut de la lemmatisation.

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

Forme	Freq. ↓	Types
developed	355	ver
rich	244	adj
romantic	204	nom
beautiful	166	adj
advanced	96	ver
civilized	78	ver
small	66	adj
freedom	65	nom
open	65	adj
classical	64	nom
high	62	adj
good	59	adj
elegant	57	nr
civilization	51	nom
environment	47	nom
clean	41	adj
welfare	41	nom
countries	40	nom
democracy	39	nom
leisure	36	nom

- 1^{er} onglet • Résumé = description générale du corpus (nombre de textes, d'occurrences, de formes...)
- 2^{ème} onglet • Formes actives = liste des formes/mots actifs (avec leur catégorie grammaticale) par fréquences décroissantes.
- 3^{ème} onglet • Formes supplémentaires = liste des formes/mots supplémentaires par fréquences décroissantes
- 4^{ème} onglet • Total = ensemble des mots par fréquences décroissantes
- 5^{ème} onglet • Hapax = mots du corpus présents une seule fois

Sur chaque forme/mot

Clic droit → **Formes associées** permet de **visualiser les regroupements** (lemmatisation)

Clic droit → **Concordancier** affiche le contexte d'utilisation du mot dans le corpus

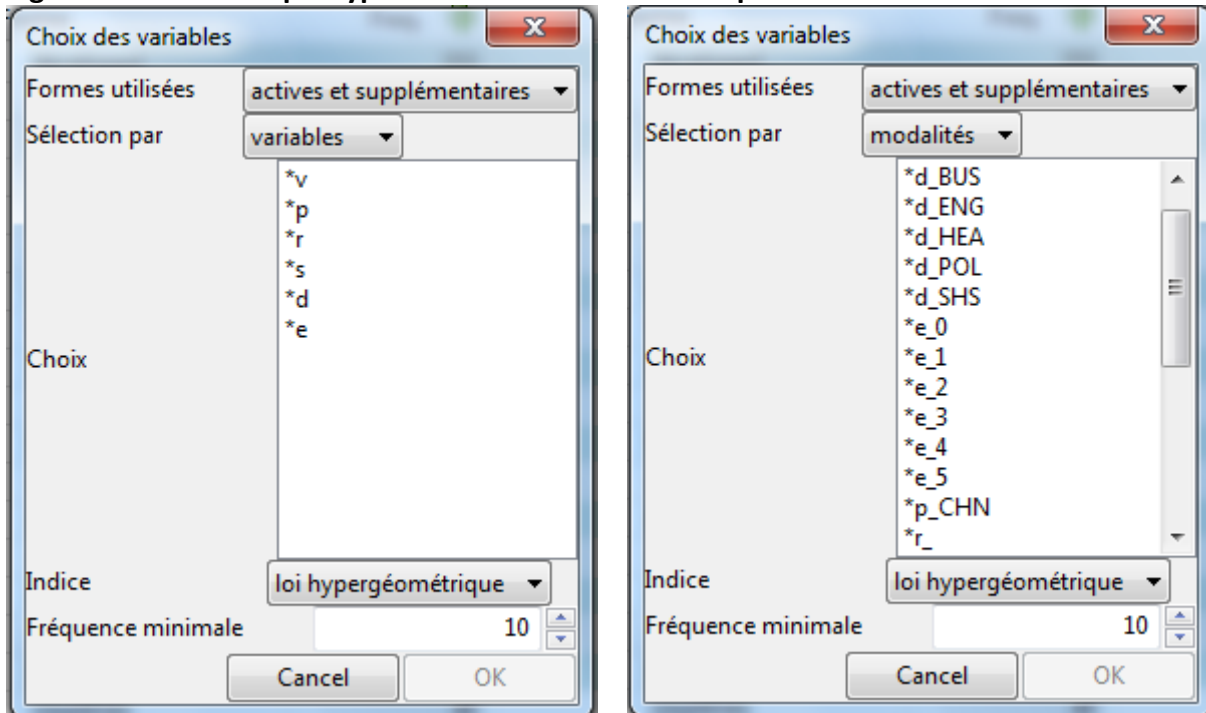
Fichiers générés de le dossier « nomducorpus_Stat_1 »:

- 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.
→ 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.

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



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

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).

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

EBM_iram_CHN_n_stat_1 Spécificités - EBM_iram_CHN_n_spec_1 Spécificités - EBM_iram_CHN_n_spec_2 x						
Formes	Formes banales	Types	Fréquences des formes	Fréquences des types	Fréquences relatives des	
formes	*v_BJS	*v_CAN	*v_NKG	*v_SHA	*v_WUH	
football	1.6901	-0.2606	-0.7303	0.459	-1.1509	
small	1.5102	-0.3163	-2.2662	0.3468	0.4345	
open	1.2917	-1.9916	0.537	-0.6227	0.6212	
good	1.0993	-0.3853	-0.8257	0.7514	-0.7337	
and	0.9451	0.3746	-0.2614	-1.439	0.4079	
a	0.9195	0.3988	-0.3203	-0.5027	-0.4799	
democratic	0.8527	0.3596	-0.3543	-1.2126	0.4711	
modernization	0.8424	0.6438	0.3648	-0.7006	-1.0988	
quality	0.8065	0.3335	-0.4018	-0.3678	-0.3626	
noble	0.7438	-0.3263	-0.2908	0.2195	-0.2679	
european	0.7438	-0.6736	0.6131	0.2195	-0.5821	
high	0.7332	0.3829	-0.7173	0.6176	-0.8502	
flourish	0.7096	-1.0691	-0.5521	-0.7416	1.9332	
graceful	0.6923	-0.5791	0.3136	-0.2515	-0.259	
comfortable	0.6835	-0.5568	0.3146	-1.0651	0.8381	
classical	0.6584	0.459	-1.3426	0.2894	0.3547	
union	0.6582	0.2348	0.2663	-0.2266	-0.6493	
united	0.6582	-1.4221	-0.6894	-0.2266	1.464	
capitalism	0.6291	-0.2213	0.4658	-0.771	-0.2847	
rich	0.5966	-0.941	-0.2742	0.6229	-0.3452	
fashion	0.5789	-0.4158	0.6461	0.2578	-0.8207	
population	0.5326	-0.2792	-0.2314	0.3759	-0.565	
cultural	0.5258	1.1392	-0.6658	-0.5969	-0.3626	

(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^{er} onglet • Formes (mots) : Affichage des formes spécifiques par modalité et par spécificité décroissante (ici formes spécifiques des étudiants interrogés à Pékin v_BJS)
- 2^{ème} onglet • Formes banales : Affichage des formes par effectif décroissant
- 3^{ème} onglet • Types (adjectif, pronom...) : catégories grammaticales
- 4^{ème} onglet • Effectifs par formes/mots
- 5^{ème} onglet • Effectifs par types de catégories grammaticales
- 6^{ème} onglet • Effectifs relatifs des formes/mots
- 7^{ème} onglet • Effectifs relatifs par type grammatical de mot
- 8^{ème} onglet • AFC (analyse factorielle des correspondances) sur un tableau lexical agrégé (TLA)

Sur chaque mot :

Clic droit → **formes associées** permet de **visualiser les regroupements (lemmatisation)**

Clic droit → **concordancier** affiche le contexte d'utilisation du mot dans le corpus

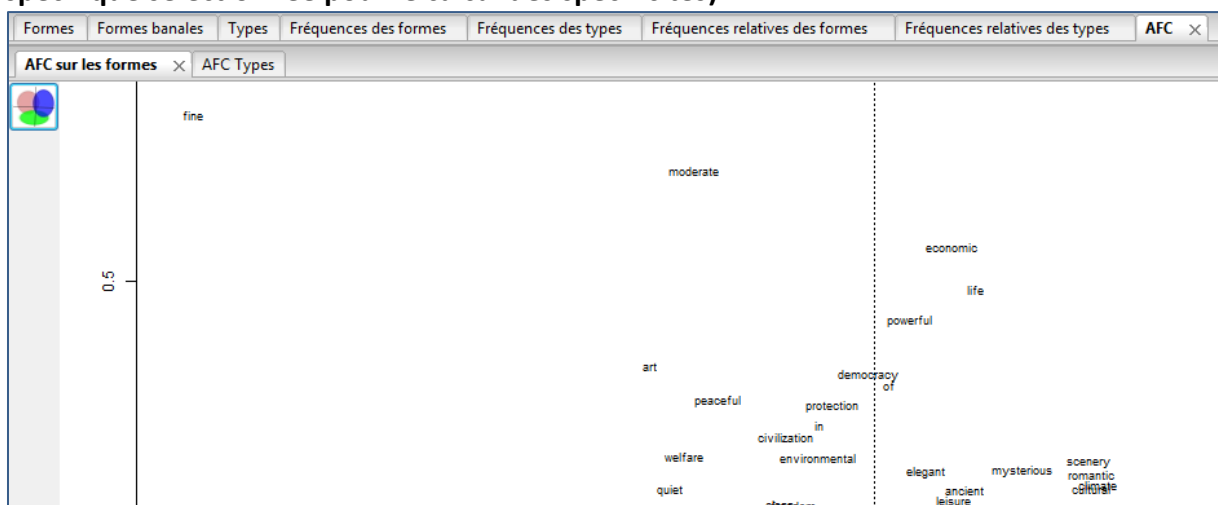
Clic droit → **graphique** affiche un graphique représentant le sur/sous emploi du mot

Clic droit → **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)


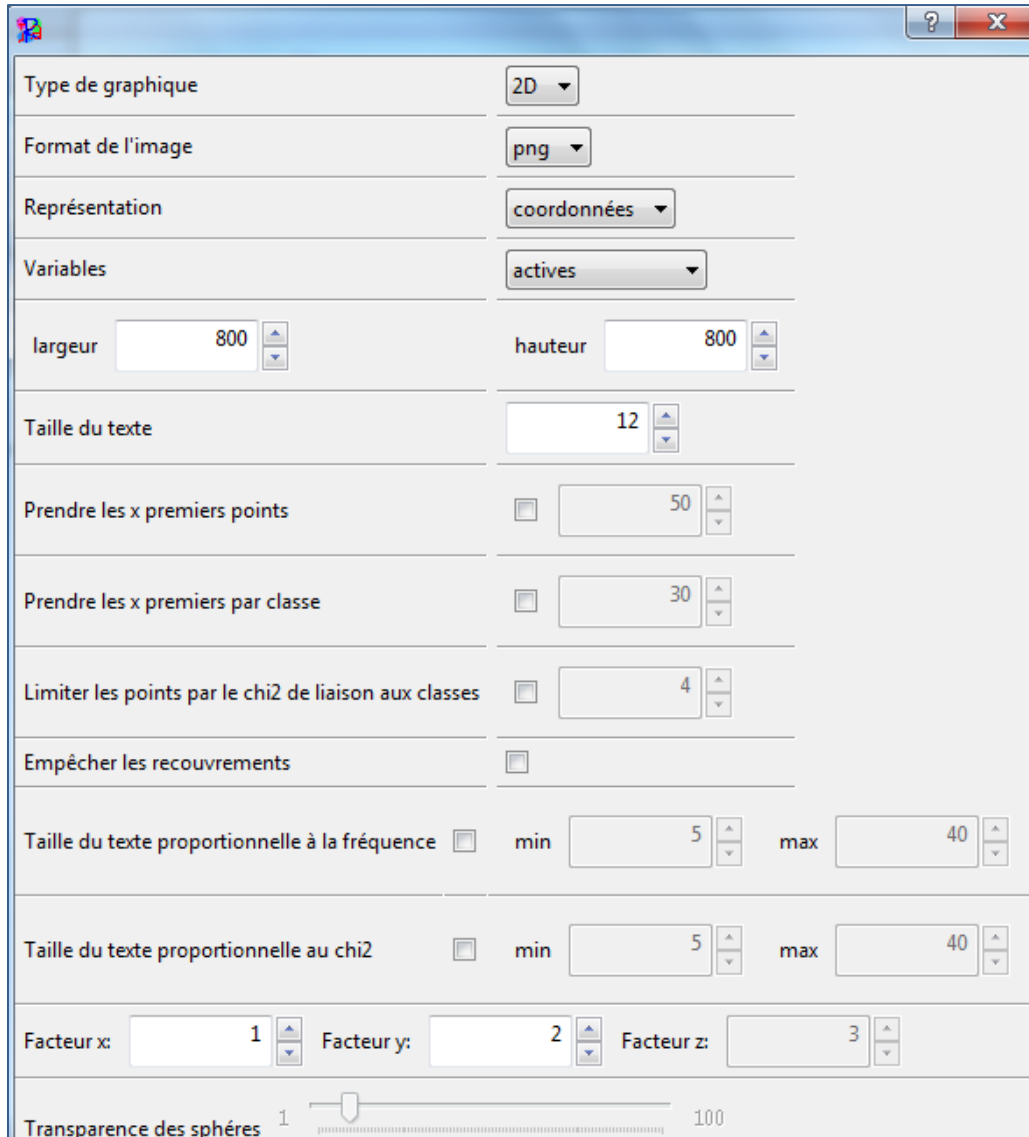
En cliquant sur ce symbole  on peut paramétrer le graphique des plans factoriels

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 Inkscape⁵ 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.

⁵ <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 ; Corrélacion 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)
→ *Equivalent de l'onglet « Effectifs formes » dans le logiciel*
- **tablespectf.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

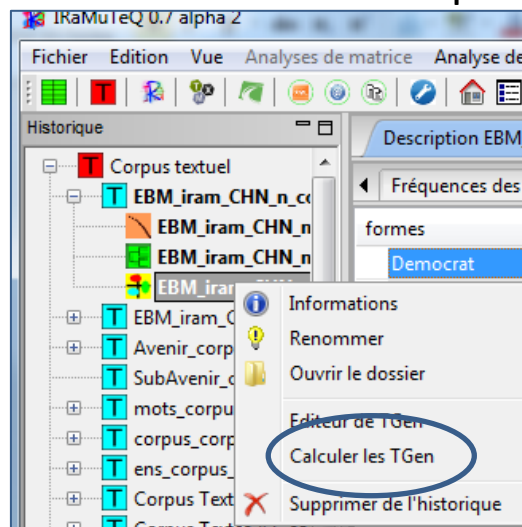


Figure 16 - Accès au menu *Faire un TGen* à partir du menu Spécificités et AFC

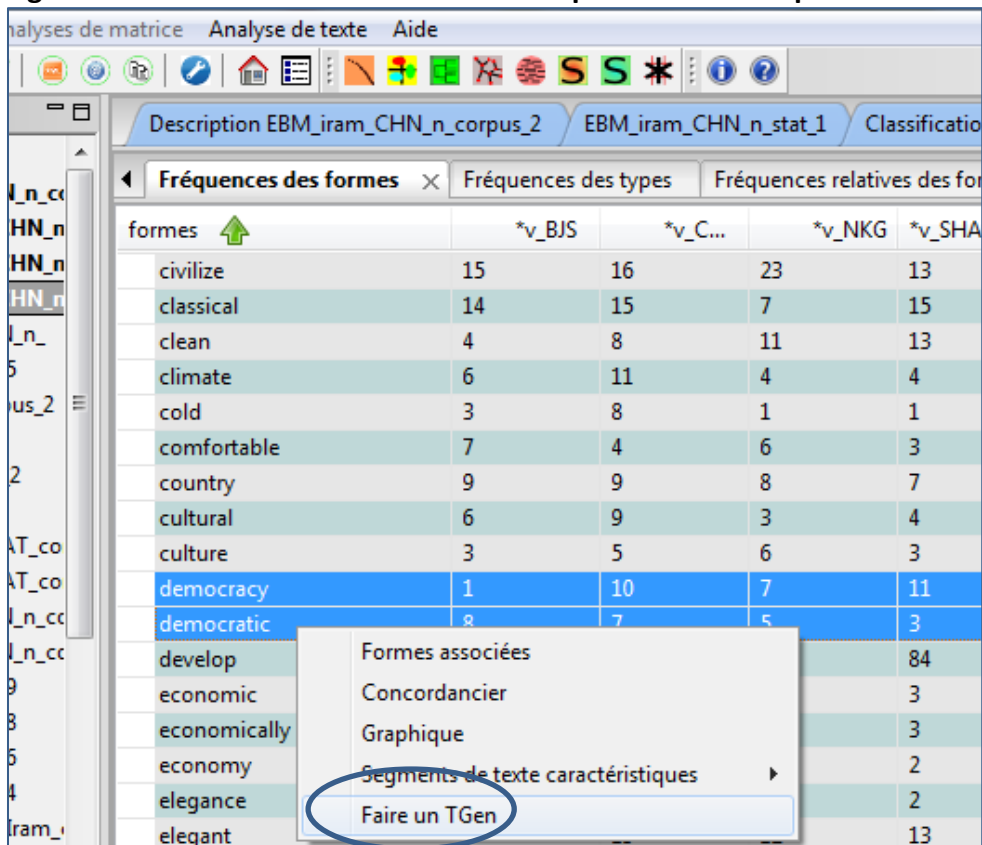
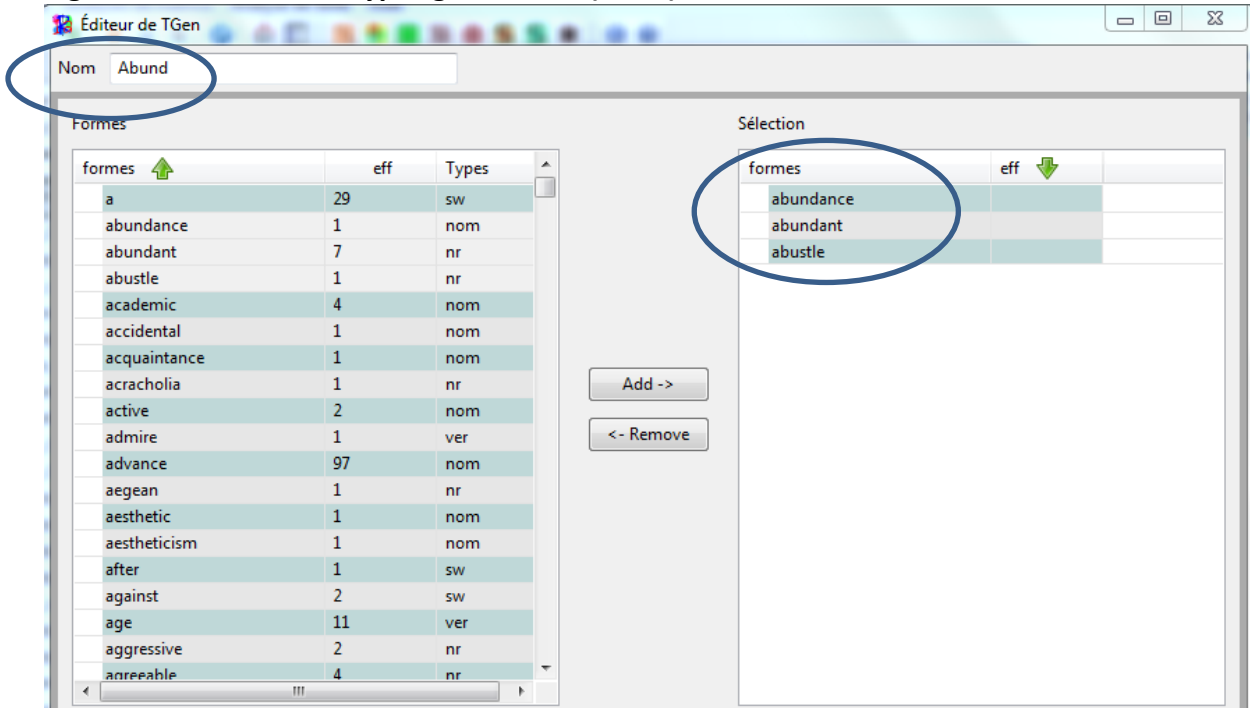


Figure 17 - Création de types généralisés (TGen)



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_n_corpus_2		EBM_iram_CHN_n_stat_1		Classification - EBM_iram_CHN_n_corpus_2		Spécificités - EBM_iram_CHN_n_spec_1	
Fréquences des formes		Fréquences des types		Fréquences relatives des formes		Fréquences relatives des types	
formes		*v_BJS	*v_CAN	*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	

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

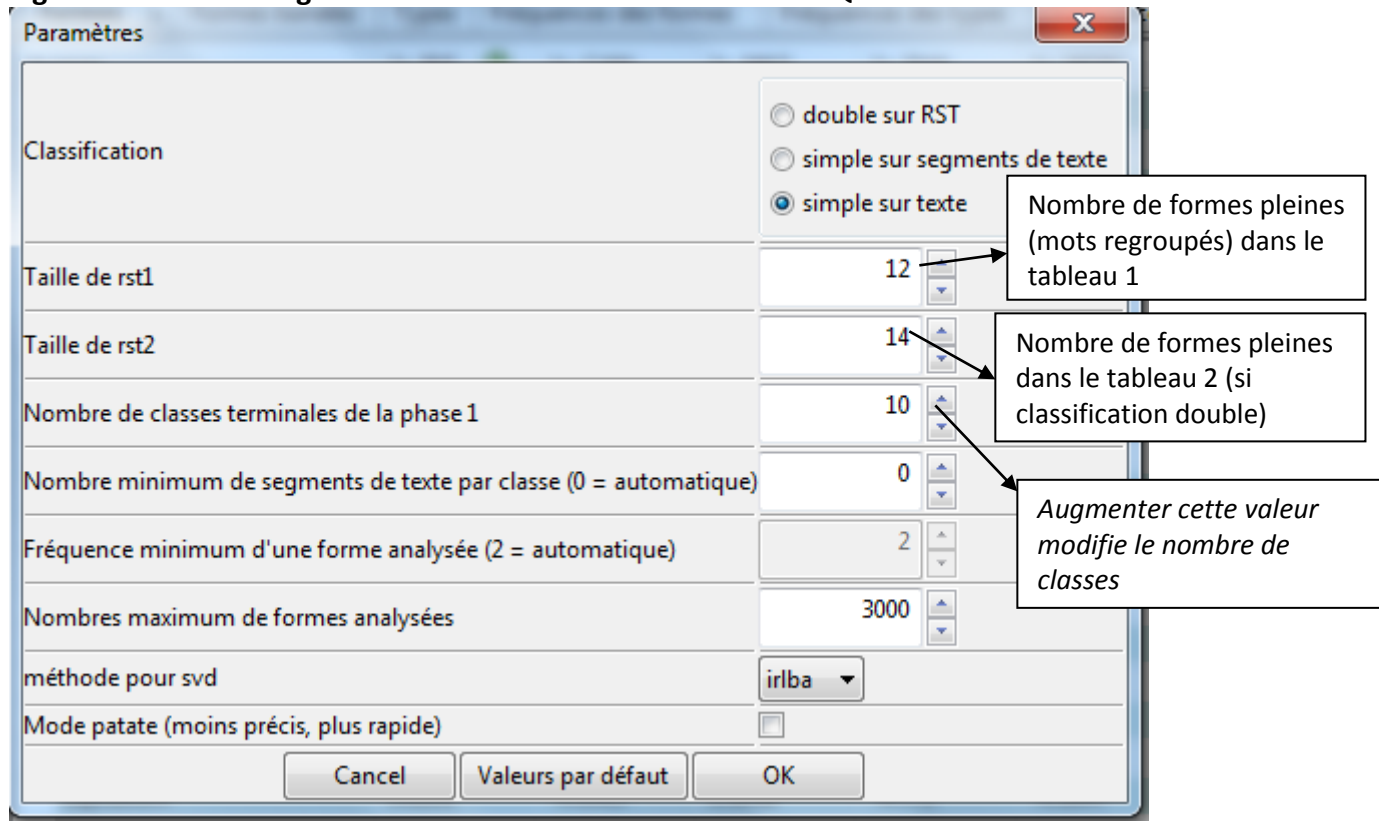
(EuroBroadMap 2009)

4 / Classification

Méthode Reinert

Implémentation de la méthode de classification « Alceste » de Max Reinert⁶ (Figure 19).

Figure 19 - Paramétrage de la méthode Reinert dans IRaMuTeQ



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.

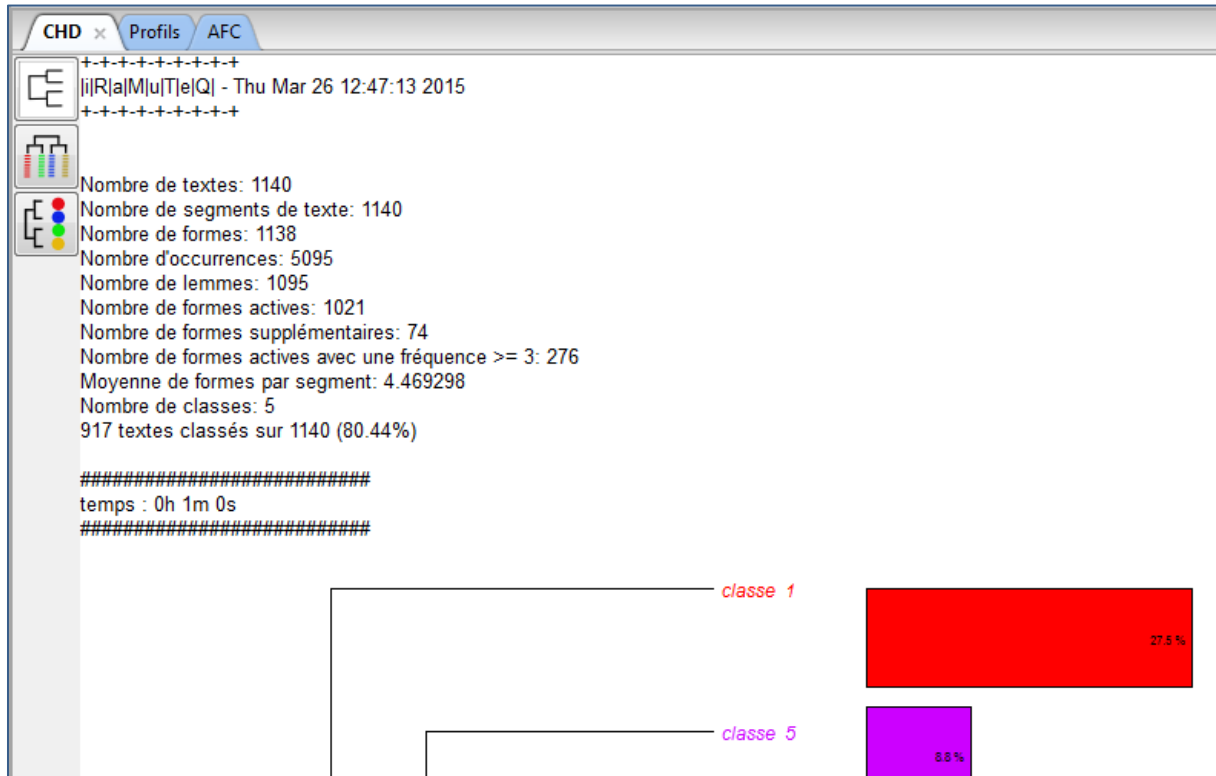
Pour afficher (et imprimer) le rapport d'analyse (équivalent du contenu de l'onglet « Profil ») faire un clic droit sur le nom de l'analyse correspondante (*nomcorpus_alceste_n*) de la fenêtre « Navigateur ».



⁶ 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



Cliquer sur les boutons   permet d'afficher les mots spécifiques des classes pour 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)

Figure 21 - 2e onglet : Profils

num	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	53	59	89.83	123.0	adj	high	< 0,0001
1	50	57	87.72	110.66	adj	small	< 0,0001
2	46	50	92.0	110.46	adj	good	< 0,0001
3	37	39	94.87	92.82	nom	welfare	< 0,0001
4	41	46	89.13	92.36	nom	environment	< 0,0001
5	27	28	96.43	68.89	nom	country	< 0,0001
6	26	28	92.86	68.89	nom	country	< 0,0001
7	24	27	88.89	68.89	nom	country	< 0,0001
8	23	26	88.46	68.89	nom	country	< 0,0001
9	15	15	100	68.89	nom	country	< 0,0001
10	15	16	93.75	68.89	nom	country	< 0,0001
11	13	13	100	68.89	nom	country	< 0,0001
67	30	49	61.22	68.89	nom	country	< 0,0001
12	15	19	78.95	68.89	nom	country	< 0,0001
13	12	14	85.71	68.89	nom	country	< 0,0001
15	9	9	100	68.89	nom	country	< 0,0001
14	9	9	100	68.89	nom	country	< 0,0001
68	13	16	81.25	68.89	nom	country	< 0,0001
69	14	18	77.78	68.89	nom	country	< 0,0001
16	10	11	90.91	68.89	nom	country	< 0,0001

→ 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).

Onglet AFC : 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).

Onglet Facteurs : On trouve les valeurs propres, pourcentage et pourcentage cumulés issus du calcul de l'AFC ;

Ici les fichiers générés sont sauvegardés dans un répertoire : **nomcorpus_alceste_1** :

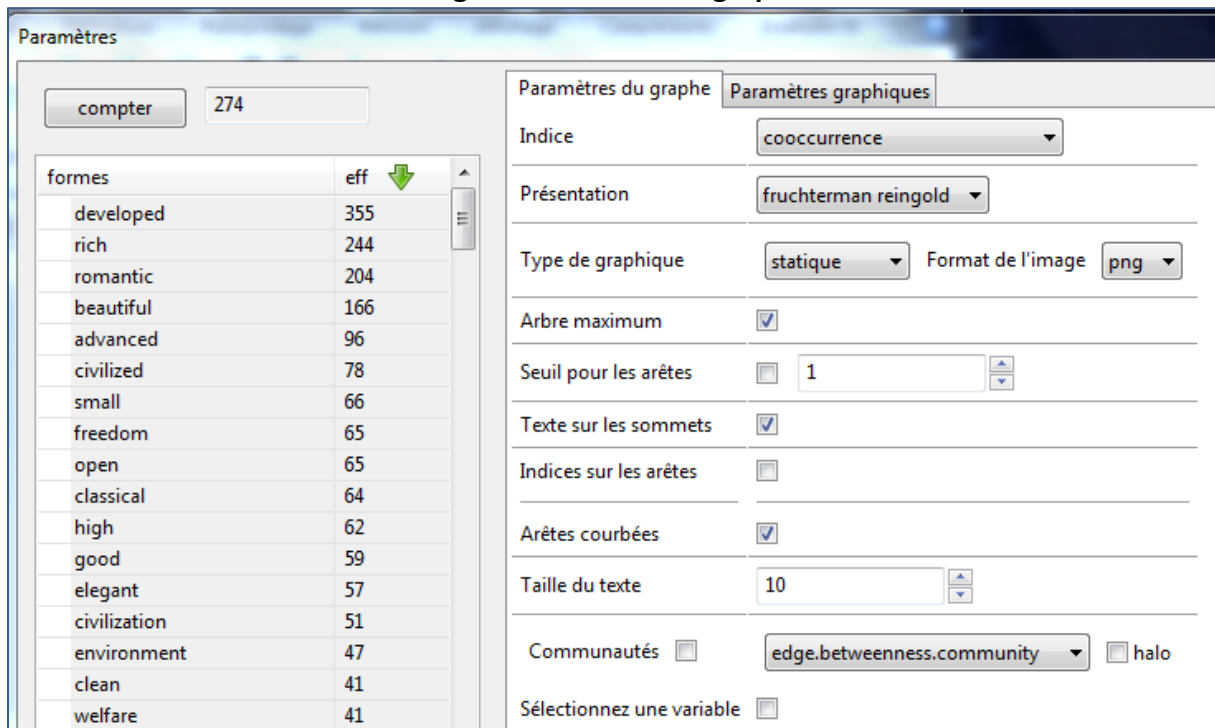
Autres fichiers qu'on peut exporter (par clic droit sur le nom du répertoire dans la fenêtre historique):

- 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

<p>**** *n_241 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc1</p> <p>clean fashionable healthy civilized</p>	
<p>**** *n_242 *p_CHN *v_BJS *s_F *d_ART *e_0 *r_Inc3</p> <p>developed economy beauteous environment linguistic diversity</p>	
<p>**** *n_244 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc2</p> <p>small area small population good environment beautiful scenery</p>	
<p>**** *n_245 *p_CHN *v_BJS *s_M *d_ART *e_2 *r_Inc3</p> <p>gleichschaltung contradiction civilized bright future</p>	EuroBroadMap, 2009

**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).

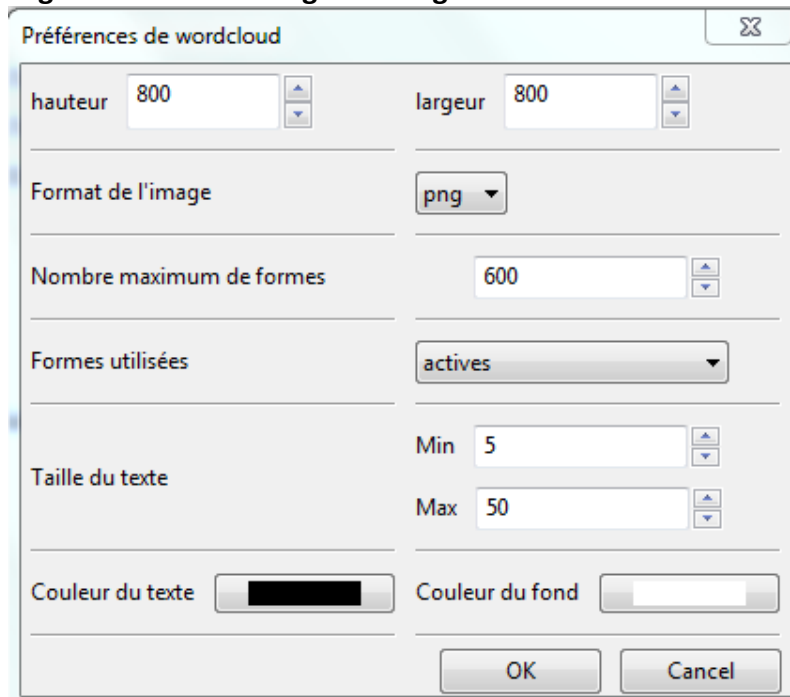
Les formes/mot mots les plus centraux sont détectés à partir du calcul de leur centralité d'intermédierité. Ces mots servent d'intermédiaires pour relier (au sens de la cooccurrence) un grand nombre d'autres mots entre eux.

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).

6 / Nuage de mots

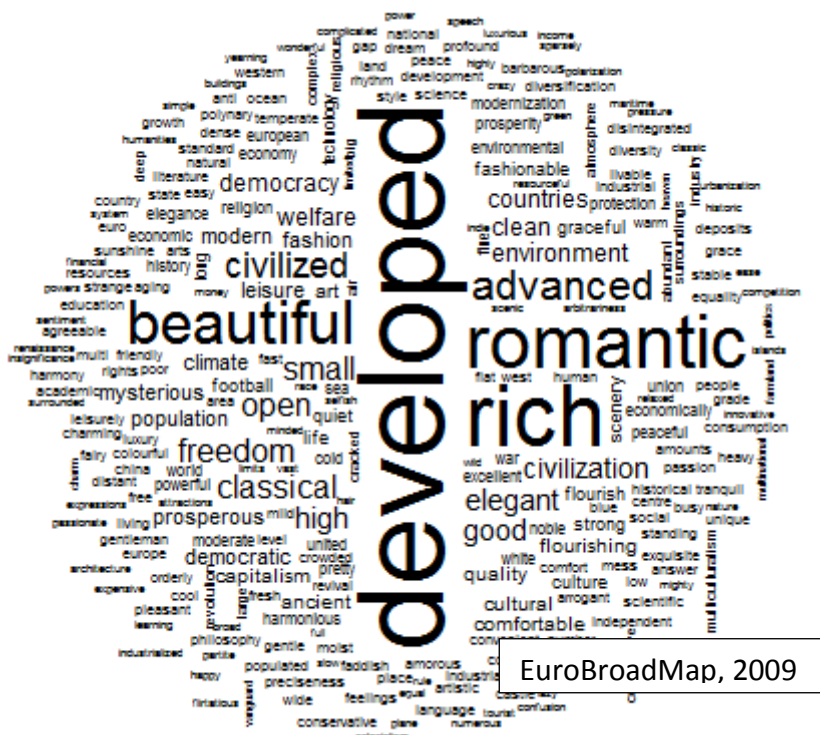
Figure 29 - Paramétrage du nuage de mots dans IRaMuTeQ



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

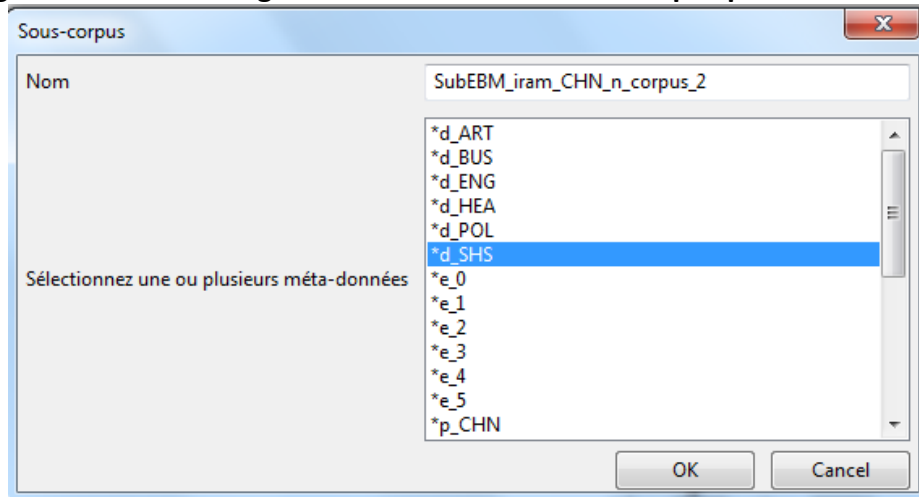


7 / Création de sous corpus

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).

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)

Figure 34 - Bilan lexical du sous-corpus créé

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	D:\ADTextuelles\Test_Iramuteq\Tests\EBM\EBM_iram_CHN_n_corpus_2\SubEBM_iram_CHN_n_corpus_2_1
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

Avril 2015

Dans la partie « Classification », nous avons vu qu'il est également possible de créer un sous-corpus 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

- http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf
- 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
- <http://www.eurobroadmap.eu/>
- Brennetot A., Emsellem K., Guérin-Pace F et Garnier B, « Dire l'Europe à travers le monde », *Cybergeo : European Journal of Geography* [<http://cybergeo.revues.org/25684>]
- 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 <i>Faire un TGen</i> 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