Novembre 2013

Utilisation d'un outil de statistiques textuelles[1]

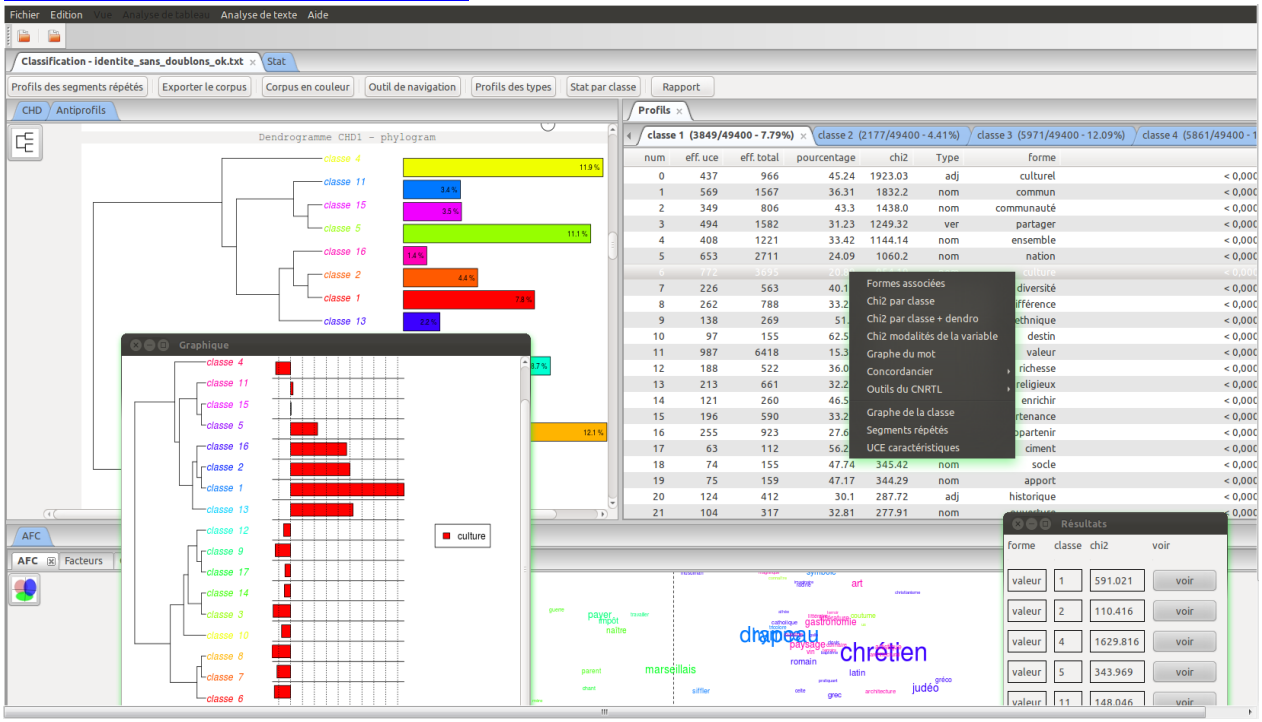
IRaMuteQ 0.6

Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires

Elodie Baril et Bénédicte Garnier
Institut National d'Etudes Démographiques

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/).
Nous traitons les réponses des étudiants chinois à une question ouverte posée comme suit : « Quels sont les mots que vous associez le plus à l'« Europe » ? Choisissez 5 mots au maximum ».

[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.

Installer IRaMuTeQ

Le logiciel est gratuit, il faut le télécharger à partir du site. Il nécessite également l'installation d'une version récente du logiciel R (et de préférence par la dernière).

Tutoriel

Une documentation sur le formatage des corpus texte est disponible sur le site du logiciel.

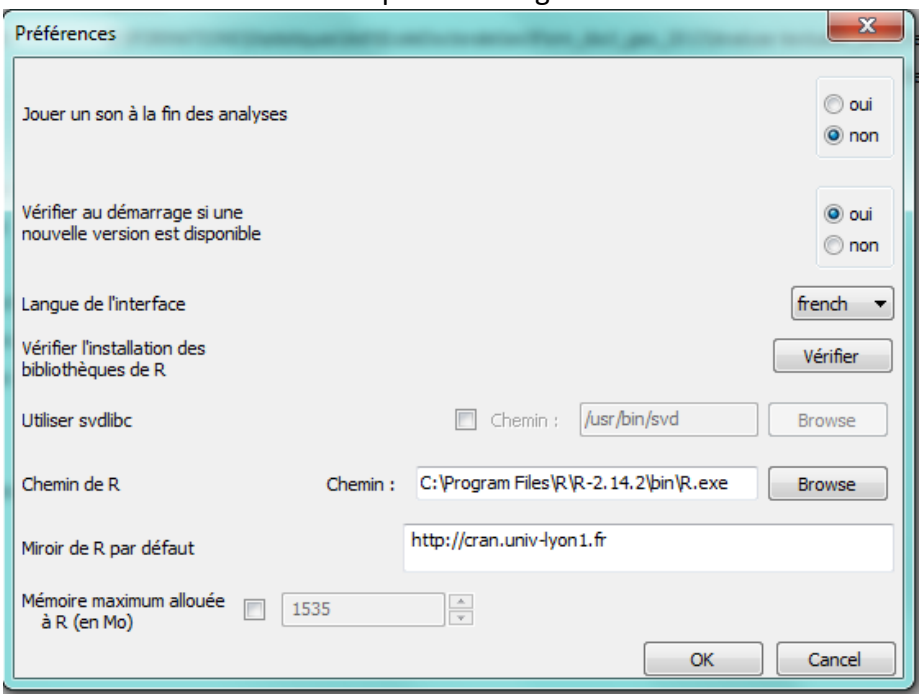
Les résultats

Les résultats des différentes opérations de l'analyse textuelle sont sauvegardés au fur et à mesure de l'exécution dans des sous-répertoires par type d'analyse. Les analyses sont sauvegardées dans un fichier (.ira). Les calculs sont également sauvegardés (au format csv) dans les sous-répertoires.

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

Description des menus d'IRaMuTeQ

Edition → Préférence : Options du logiciel



1ère étape : Importer le fichier à analyser

Menu Fichier ouvrir …

Dans ce document, nous ne traitons pas les menus *ouvrir une matrice* (importation d'un tableau de données comportant des valeurs numériques) et *importer de 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).

Nous choisissons le menu ouvrir un corpus

Le fichier à analyser est un fichier texte qui respecte la mise en forme avec ligne étoilée « Alceste »[2]. Les textes à analyser sont très courts et on dispose de caractéristiques sur ces textes (pays de naissance des enquêtés, ville d'enquête, sexe, etc.)

[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 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. Attention, il est recommandé de mettre au minimum 2 variables étoilées pour que toutes les analyses fonctionnent, notamment la classification.

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
```

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.

Première analyse du logiciel : création du lexique

→ Ouverture de la fenêtre de paramétrage avant le lancement de l'analyse.
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 pour la reconnaissance des catégories de mots (dans la lemmatisation).
Il est possible de changer le répertoire destination des résultats **Répertoire en sortie**.

Le **marqueur de textes** correspond au séparateur entre unités d'analyses (ici les 1140 réponses des étudiants chinois). Nous avons les identifiants des questionnaires (sur 4 caractères).

Les **dictionnaires d'expression** repèrent des expressions courantes comme « aujourd'hui » ou « grand-père » (en français)

Construire des segments de texte : Si le corpus est long (cas d'entretiens par exemple) IRaMuteQ propose de découper les textes en unités plus petites. Ce découpage peut se faire en fonction d'un nombre d'occurrences (cas par défaut) ou de caractères ou de paragraphes.

Possibilité aussi de découper le corpus (par variable ou par modalité).

Figure 1 : Indexation du corpus

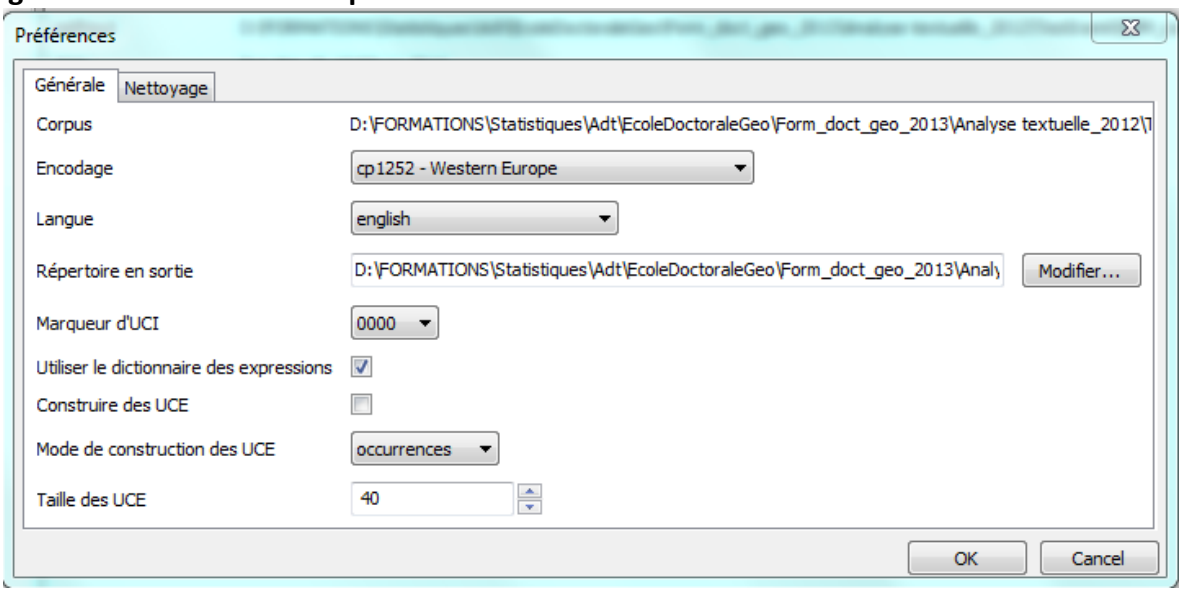
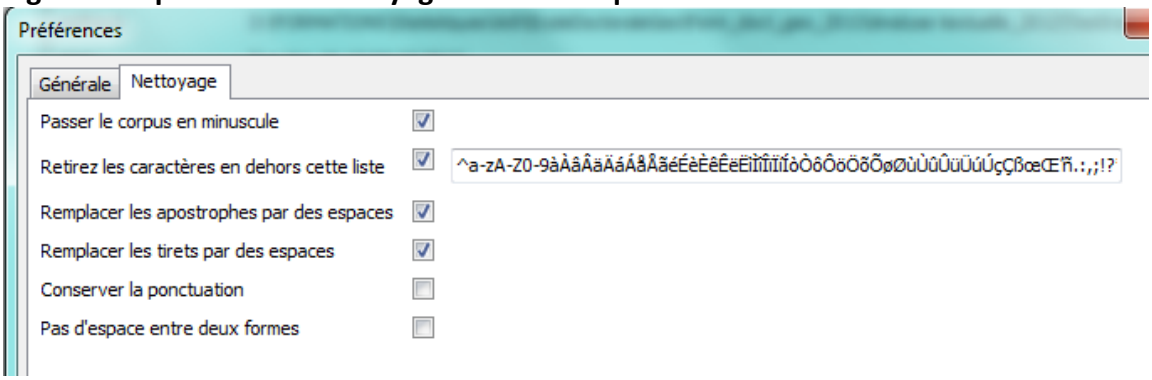


Figure 2 : Options du « Nettoyage » automatique du fichier



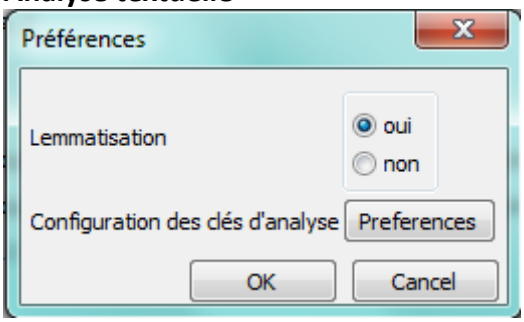
Une fois le corpus préparé, IRaMuteQ affiche un bilan

Menu Analyse de Texte

1) Analyse de Texte→ Statistiques textuelles

Lemmatisation et paramétrage des catégories de mots à prendre en compte dans les calculs.

Analyse textuelle

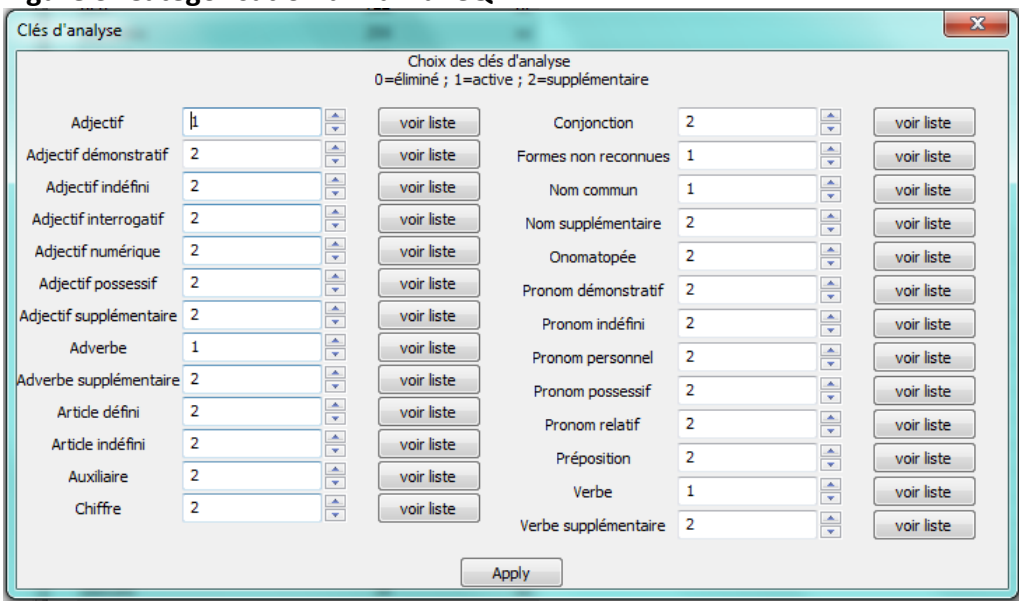


→ Équivalent des **clés catégorielles** d'Alceste

Par défaut, le logiciel fait une lemmatisation à l'aide de ses dictionnaires[3]
Il reconnaît les catégories grammaticales des mots et les expressions. Selon le type, il les traitera en élément *actif* ou *supplémentaire* (Garnier, Guérin-Pace, 2010).

Configuration des clés d'analyse → Préférences (permet de modifier les clés d'analyse par catégories).

Figure 3: Catégorisation d'IRaMuTeQ



- 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.

[3] Dictionnaires anglais, allemands, italiens, espagnols, portugais (certains sont encore expérimentaux), dictionnaires minimalistes pour le suédois et le grecs.

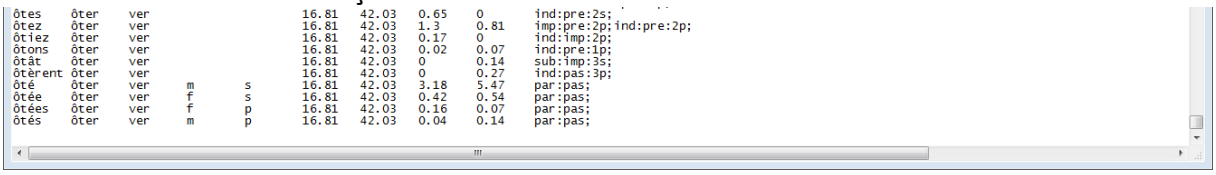
Modifier le(s) dictionnaire(s)

- 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



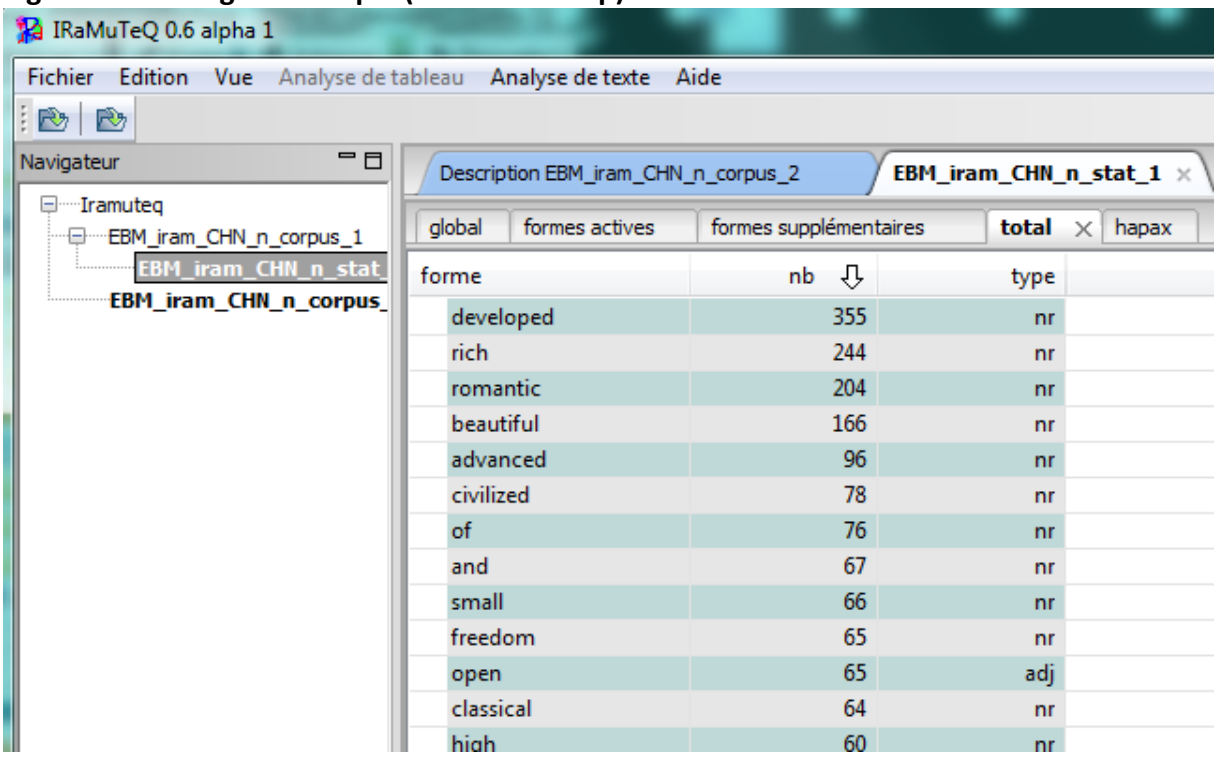
ôtes	ôter	ver			16.81	42.03	0.65	0	ind:pre:2s;
ôtez	ôter	ver			16.81	42.03	1.3	0.81	imp:pre:2p;ind:pre:2p;
ôtiez	ôter	ver			16.81	42.03	0.17	0	ind:imp:2p;
ôtons	ôter	ver			16.81	42.03	0.02	0.07	ind:pre:1p;
ôtât	ôter	ver			16.81	42.03	0	0.14	sub:imp:3s;
ôtèrent	ôter	ver			16.81	42.03	0	0.27	ind:pas:3p;
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)

Répertoire(ou dossier) généré par IRaMuTeQ : *nomdufichier texte*_stat1.

Pour toutes les analyses, un clic droit sur une analyse ou un corpus permet d'afficher les options utilisées pour le traitement.
Il est également possible d'exporter le dictionnaire d'un corpus ou le dictionnaire des lemmes partir d'une analyse statistique.

Figure 4: Affichage du lexique (EuroBroadMap)



1er onglet : Global = description du corpus (nombre de textes, occurrence, formes…)
2ème onglet : formes_actives = liste des formes/mots actifs par fréquences décroissantes
3ème onglet : formes_supplémentaires = liste des formes/mots supplémentaires par fréquences décroissantes
4ème onglet : total = ensemble des mots par fréquences décroissantes
5ème onglet : hapax = mots du corpus présents une seule fois

Sur chaque mot
clic droit → formes associées permet **visualiser les regroupements (lemmatisation)**
clic droit → concordancier = affiche le contexte d'utilisation du mot dans la corpus

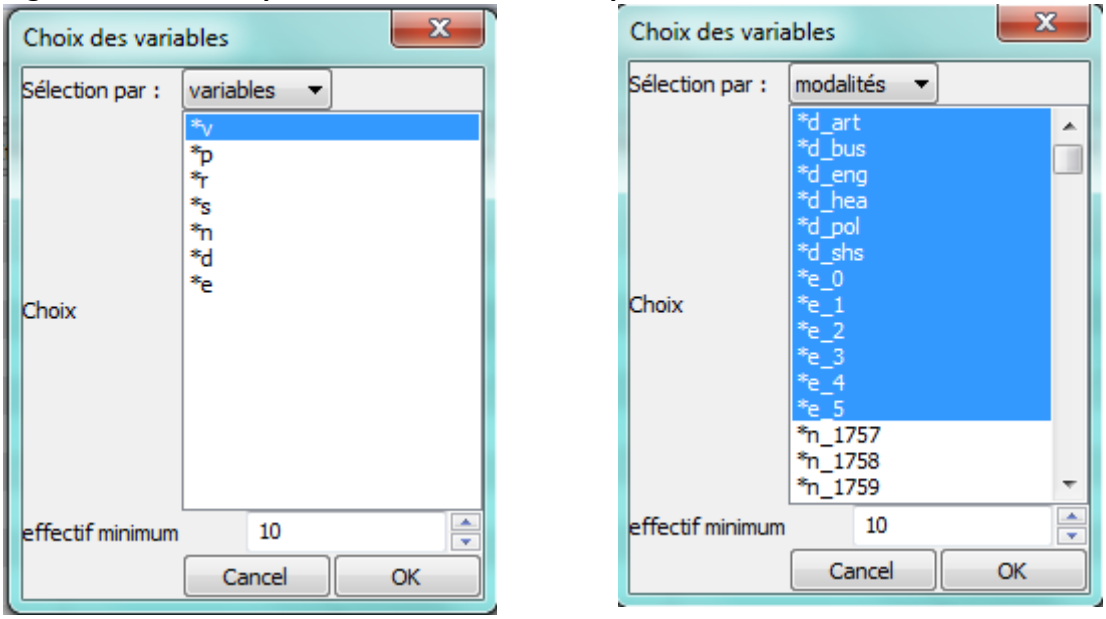
Fichiers générés de le dossier « nomducorpus_Stat_1 »:

- analyse.ira : Fichier permettant d'ouvrir l'analyse déjà faite dans le logiciel.
- corpus (txt) : Contient toutes les unités statistiques (réponses) en lignes
- formes_actives (csv) : 3 colonnes : chaque mots en ligne que le logiciel prend en compte ; leur fréquence, type de mots.
- 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).
- hapax (csv) : mots ayant une fréquence de 1.
- glob (txt) : fichier Global : nombre de textes : 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 texte : 4.47
- total (csv) : Tous les mots, fréquences décroissante à partir de 2 citations.
- formes_formes (csv) : mots en 1 ère colonnes, fréquences en 2 ème colonne, type en 3 ème colonne et le numéro affecté au mot par le logiciel (numérisation) en 4ème
→ Permet de visualiser les mots non lemmatisés et leur catégorie.
- graphique : fréquences en ordonnée, rangs en abscisse

2) Analyse de Texte→ Spécificités et AFC

Cette partie calcule les mots spécifiques par sous-catégories et réalise une Analyse Factorielle sur un tableau lexical agrégé (TLA)
Choix des variables pour calculer les spécificités et construire le tableau lexical

Figure 5 : Sélection par variables ou sélection par modalités



En sélectionnant par variables on ne peut choisir qu'une variable à la fois (mais IRaMuTeQ ne fait pas d'AFC avec une variable qui a trop peu de modalités).
En revanche, en faisant sélection par modalités, on peut choisir toutes les variables intéressantes et retirer les modalités rares (peu d'individus).

Figure 6 : Spécificités du corpus EuroBroadMap (étudiants chinois)

Fichier Edition Vue Analyse de tableau Analyse de texte Aide

Spécificités

formes | Types | Effectifs formes | Effectifs Type | Effectifs relatifs formes | Effectifs relatifs Type | AFC

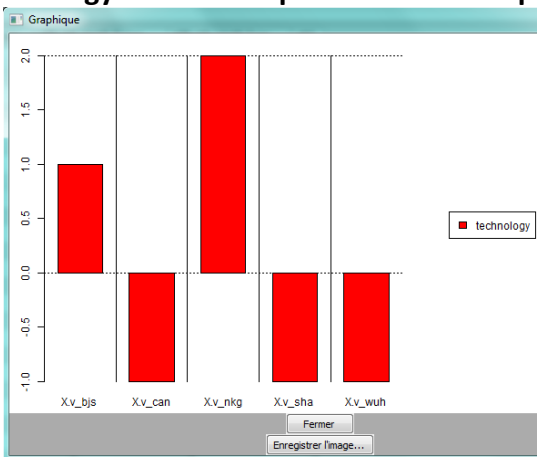
	X.v_bjs	X.v_can	X.v_nkg	X.v_sha	X.v_wuh
small	2	-1	-2	1	1
open	2	-2	1	-1	1
football	2	-1	-1	1	-2
white	1	-1	1	1	-1
war	1	-2	2	1	-2
unite	1	-2	-1	-1	2
union	1	1	1	-1	-1
technology	1	-1	2	-1	-1
sea	1	2	-2	-1	1
scenery	1	1	-1	-1	1
rich	1	-1	1	1	-1

1er onglet : Formes (mots): mots spécifiques

2ème onglet : Types (adj, pronom…) : catégories grammaticales
3ème onglet : Effectifs par formes/mots
4ème onglet : Effectifs par types de catégories grammaticales
5ème onglet : Effectifs relatifs des formes/mots
6ème onglet : Effectifs relatifs par type

Sur chaque mot
clic droit → formes associées permet **visualiser les regroupements (lemmatisation)**
clic droit → concordancier = affiche le contexte d'utilisation du mot dans la corpus
clic droit → graphique = affiche un graphique représentant le sur/sous emploi du mot

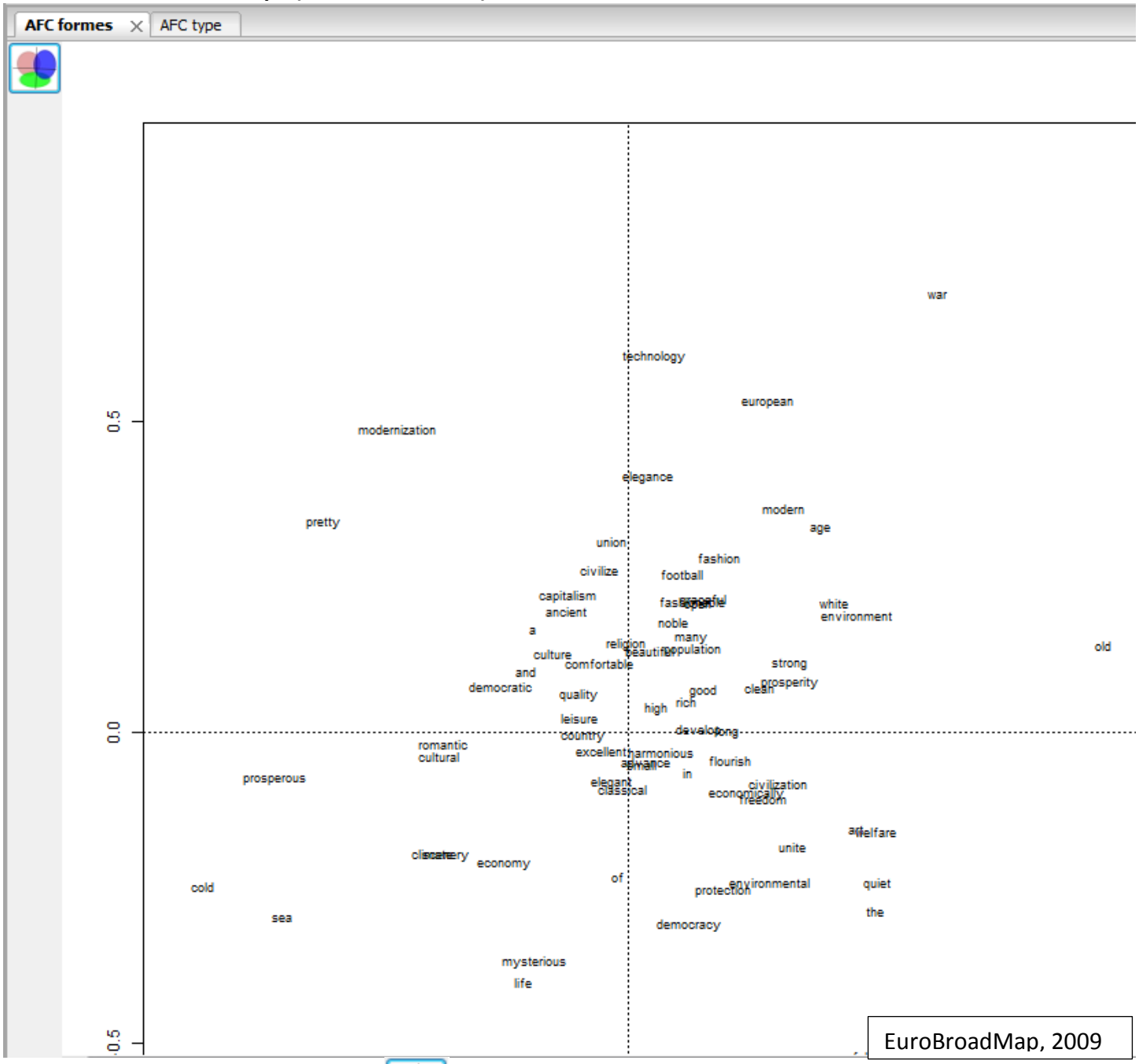
Figure 7: Emploi du mot "Technology" dans le corpus EuroBroadMap (étudiants chinois)



7ème onglet : AFC sur tableau lexical agrégé (TLA)

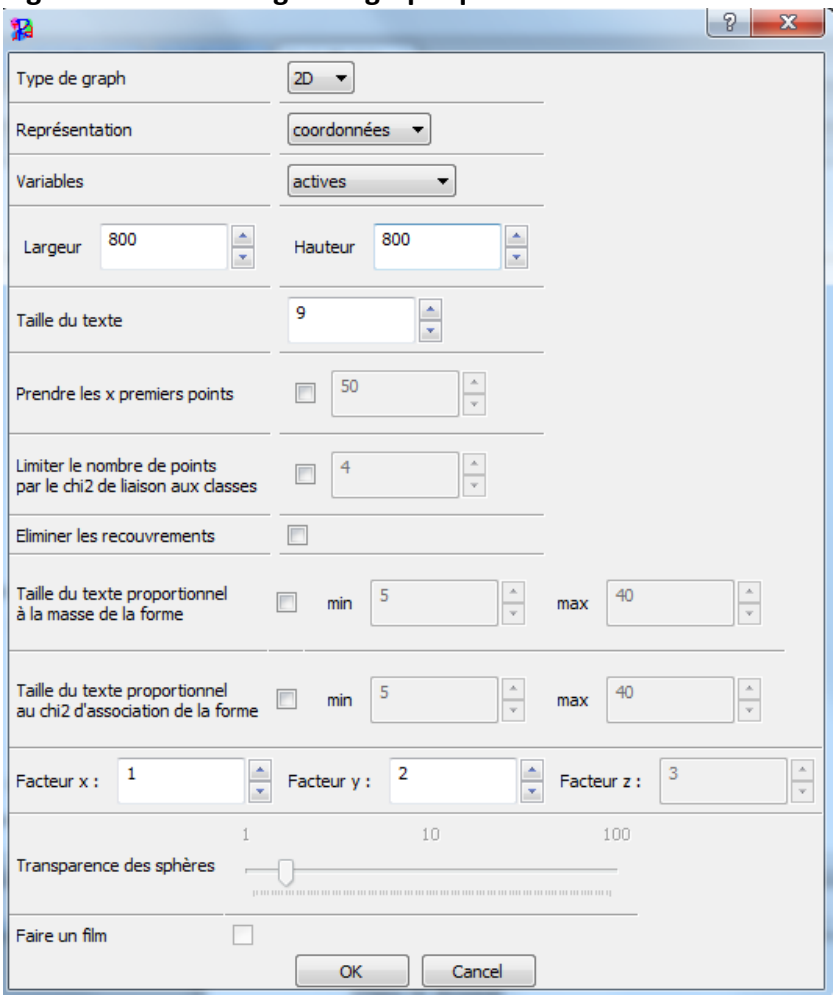
-AFC forme : génère un graphique avec tous les mots analysés et un graphique avec les variables étoilées.
-AFC type : génère un graphique avec le type des mots et un graphique avec les variables étoilées.

Figure 8 : AFC sur le Tableau Lexical Agrégé (mots et variables sélectionnées) "EuroBroadMap" (étudiants chinois)



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

Figure 9 : Paramétrage des graphiques issus d'AFC



Type de graph : choix entre 2D et 3D
Représentation : choix **entre coordonnées et corrélation**
Variables : choix entre actives, supplémentaires, étoilées, classes

Remarque : il n'est pas possible de déplacer les mots du graphique pour une meilleure visibilité. Possibilité d'imprimer ce graphique en l'ouvrant dans le dossier crée (fichier au format png). 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.

Sortie résultats du logiciel (nomcorpus_Stat_1_lexico_1) :

Sortie dossiers nomtexte_Stat_1_lexico_1 ou après une classif : nomtexte_lexico_1 (on trouve tous les calculs d'AFC (contributions, coordonnées, ….):

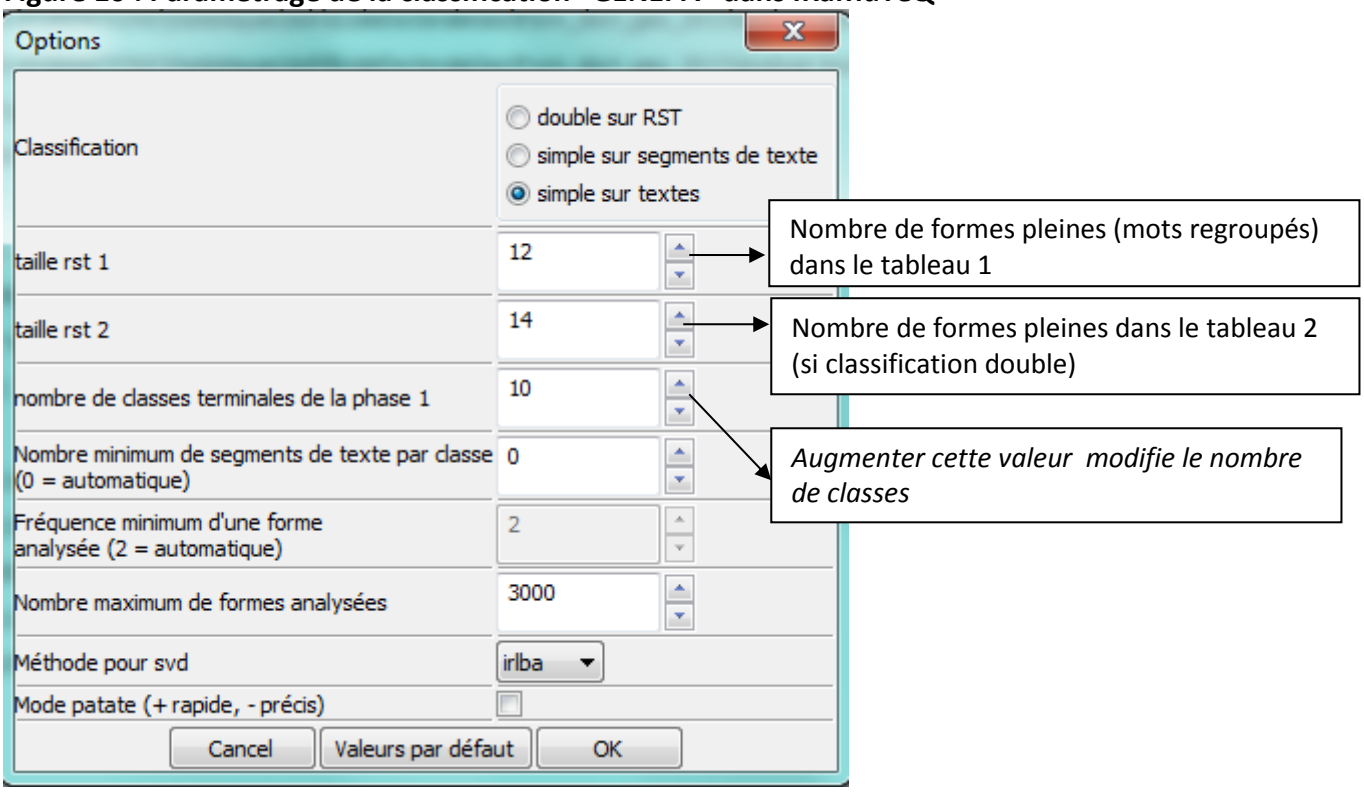
- **afcf_col.csv** : Ligne : classes
  Colonnes : Coord. facteur ; Corr. facteur 1 à 6 ; COR -facteur 1 à 6 ; CTR -facteur 1 à 6 (contribution) ; mass ; chi.distance ; inertie
- **afcf_col.png** : image / graphique des modalités actives (var étoilées)
- **afcf_facteur.csv** : Ligne : facteurs
- Colonnes : 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.db** : fichier de base de données du logiciel
- **Analyse.ira** : analyse qui peut être ouvert avec le logiciel dans « ouvrir une analyse → ouvre les onglets résultats de « spécificité et AFC ».
- **corpus.txt** : corpus de la variable textuelle sans les variables étoilées.
- **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
- **formes_formes.csv** (déjà vu): Ligne : tous les mots
  Colonnes : effectif, type de mots, chiffre attribué par le logiciel
- **formes_uces.csv**
- **tableafcm.csv :** Ligne : les mots retenus
  Colonnes : les modalités des variables étoilées sélectionnées (en effectif]
  → ***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)

3) Analyse de Texte→ Classification

Il était possible avant la version 6 d'IRaMuteQ d'utiliser la méthode « Alceste »[4]. Cette méthode s'intitule maintenant GNEPA.

Figure 10 : Paramétrage de la classification "GENEPA" dans IRaMuTeQ

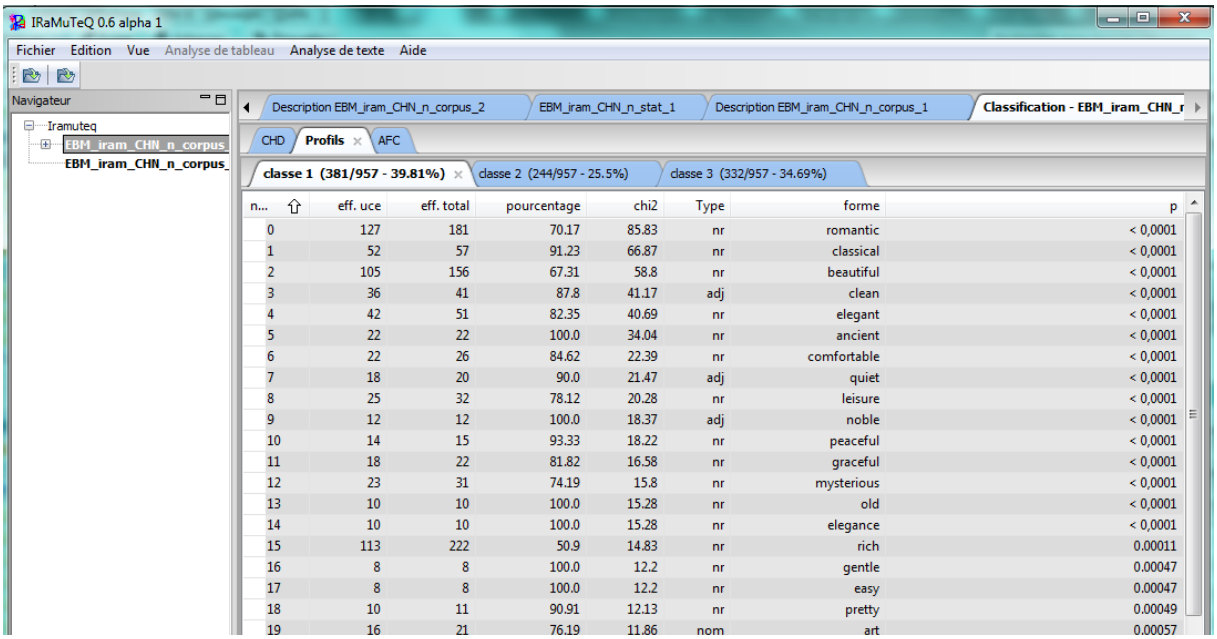


Remarque : on ne peut pas changer la « fréquence minimum d'une forme analysée » qui est en grisé. Seule la valeur du "nombre maximum de formes analysées" est prise en compte. Si le nombre total de formes actives est inférieur à cette valeur, seules les formes ayant un effectif d'au moins trois sont retenues.

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 de la fenêtre « Navigateur ».

[4] Changement de vocabulaire pour la méthode «Alceste» de Max Reinert. Le vocabulaire lié à la méthode a été remplacé. Méthode ALCESTE → Méthode GNEPA, UCI → texte, UCE → segment de texte, UC → regroupement de segment de texte (rst)

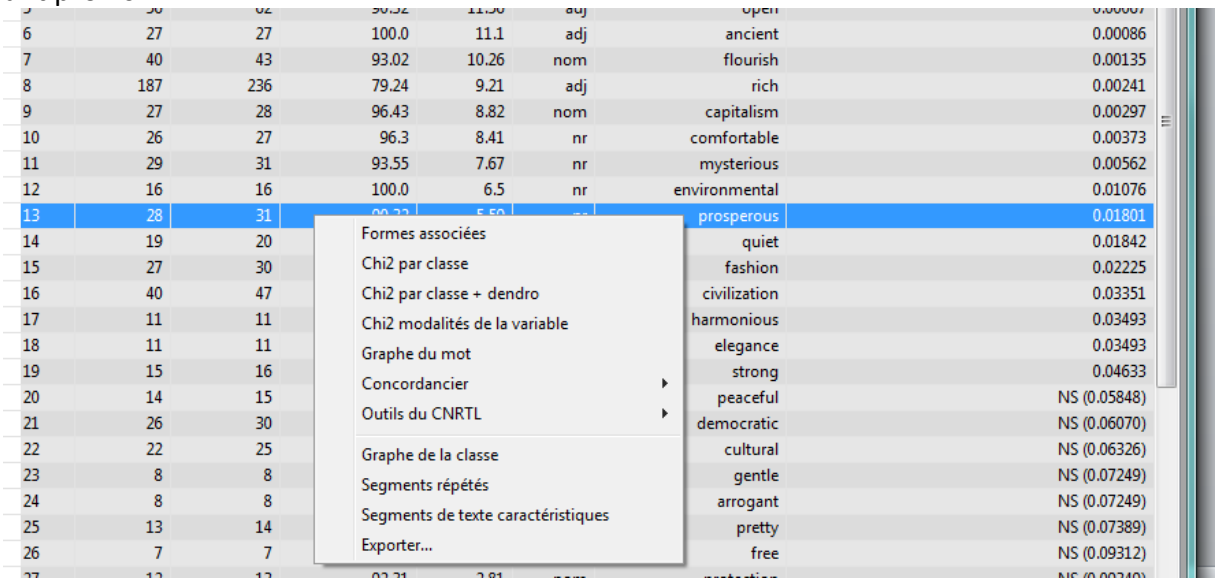
Sortie résultats de la classification :



n...	eff. uce	eff. total	pourcentage	chi2	Type	forme	p
0	127	181	70.17	85.83	nr	romantic	< 0,0001
1	52	57	91.23	66.87	nr	classical	< 0,0001
2	105	156	67.31	58.8	nr	beautiful	< 0,0001
3	36	41	87.8	41.17	adj	clean	< 0,0001
4	42	51	82.35	40.69	nr	elegant	< 0,0001
5	22	22	100.0	34.04	nr	ancient	< 0,0001
6	22	26	84.62	22.39	nr	comfortable	< 0,0001
7	18	20	90.0	21.47	adj	quiet	< 0,0001
8	25	32	78.12	20.28	nr	leisure	< 0,0001
9	12	12	100.0	18.37	adj	noble	< 0,0001
10	14	15	93.33	18.22	nr	peaceful	< 0,0001
11	18	22	81.82	16.58	nr	graceful	< 0,0001
12	23	31	74.19	15.8	nr	mysterious	< 0,0001
13	10	10	100.0	15.28	nr	old	< 0,0001
14	10	10	100.0	15.28	nr	elegance	< 0,0001
15	113	222	50.9	14.83	nr	rich	0.00011
16	8	8	100.0	12.2	nr	gentle	0.00047
17	8	8	100.0	12.2	nr	easy	0.00047
18	10	11	90.91	12.13	nr	pretty	0.00049
19	16	21	76.19	11.86	nom	art	0.00057

1er onglet : CHD : résumé (nombre de textes, segment de textes, occurrences…), dendrogramme.
2e onglet : Profils : de chaque classe

→ Plusieurs options en faisant clic droit sur les lignes (forme) des onglets profils et antiprofils :



6	27	27	100.0	11.1	adj	ancient	0.00086
7	40	43	93.02	10.26	nom	flourish	0.00135
8	187	236	79.24	9.21	adj	rich	0.00241
9	27	28	96.43	8.82	nom	capitalism	0.00297
10	26	27	96.3	8.41	nr	comfortable	0.00373
11	29	31	93.55	7.67	nr	mysterious	0.00562
12	16	16	100.0	6.5	nr	environmental	0.01076
13	28	31				prosperous	0.01801
14	19	20				quiet	0.01842
15	27	30				fashion	0.02225
16	40	47				civilization	0.03351
17	11	11				harmonious	0.03493
18	11	11				elegance	0.03493
19	15	16				strong	0.04633
20	14	15				peaceful	NS (0.05848)
21	26	30				democratic	NS (0.06070)
22	22	25				cultural	NS (0.06326)
23	8	8				gentle	NS (0.07249)
24	8	8				arrogant	NS (0.07249)
25	13	14				pretty	NS (0.07389)
26	7	7				free	NS (0.09312)

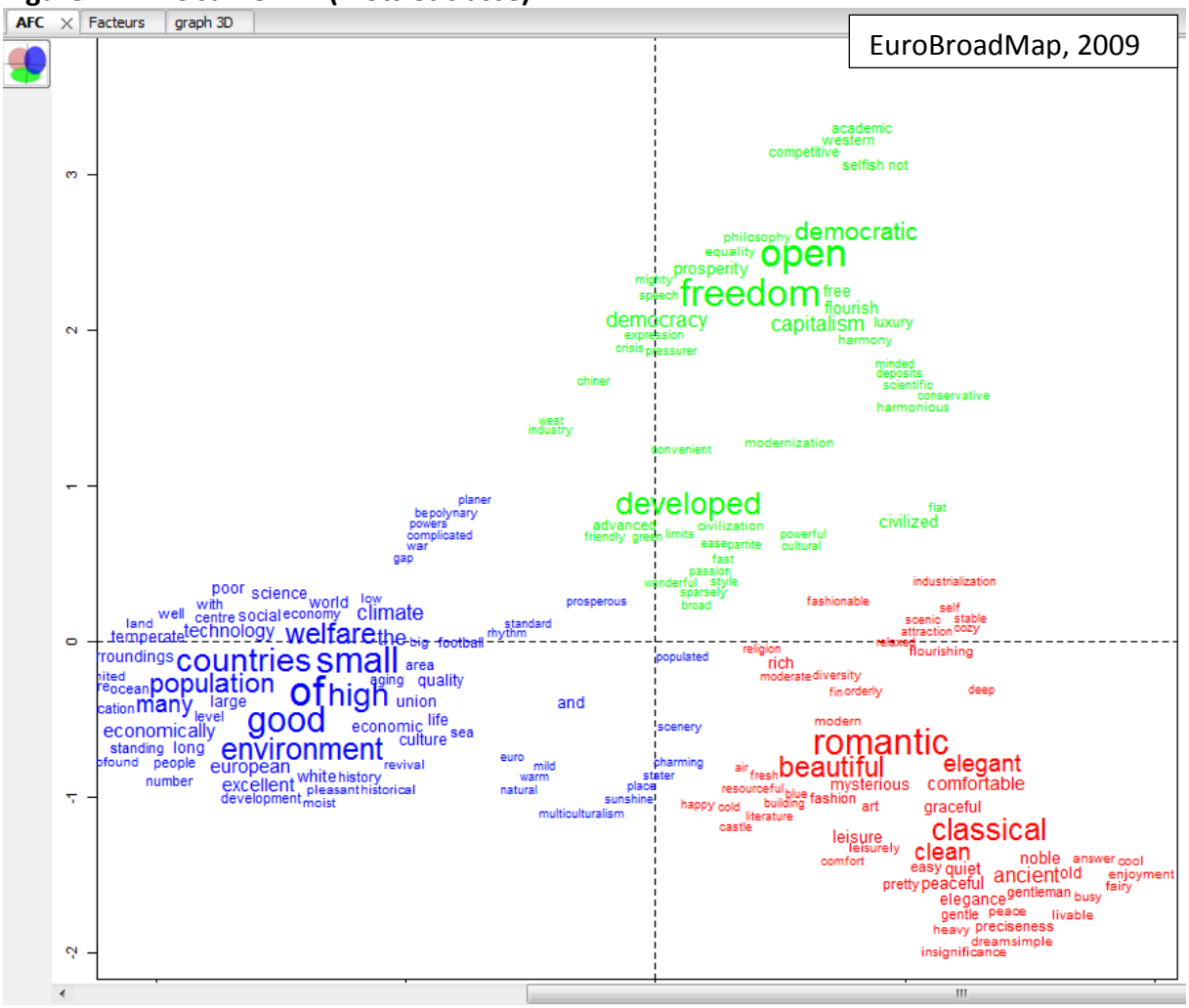
Formes associées
Chi2 par classe
Chi2 par classe + dendro
Chi2 modalités de la variable
Graphe du mot
Concordancier
Outils du CNRTL
Graphe de la classe
Segments répétés
Segments de texte caractéristiques
Exporter...

Concordancier : - dans les segments de texte de la classe
- dans les segments de texte classés
- dans toutes les segments de texte
Outils du CNRTL : renvoie sur le site du Centre National de Ressources Textuelles et Lexicales
Exporter : Permet d'exporter le corpus correspondant à cette classe

4e onglet : AFC (sur tableau croisant le lexique et la variable de classe) :

- AFC classes de différentes couleurs (selon la classe à laquelle les formes appartiennent): AFC (**coordonnées**) des variables actives (mots) ; des variables supplémentaires (chiffre, pronoms, prénoms…) ; des variables illustratives (variables étoilées) ; des classes + AFC (**corrélations**) des variables actives, supplémentaires, illustratives, des classes.
- Facteurs (valeurs propres, pourcentage, pourcentage cumulées)
- Colonnes (classes en ligne : coordonnées facteur, mass, distance du chi2, inertie)
- Lignes (individus en lignes + certains mots : coordonnées facteur, mass, distance du chi2, inertie)
- Graph 3D

Figure 11: AFC sur le TLA (mots et classe)



Possibilité d'exporter le graphique au format vectoriel (format svg) pour améliorer le rendu avec un logiciel de DAO (InkScape[5] ou Illustrator).

Autres onglets qu'on peut exporter :

- profils des segments répétés ; exporter le corpus ; **corpus en couleur** (fichier html où chaque segment de texte est associé à une couleur qui donne sa classe d'appartenance) ;

[5] http://inkscape.org/?lang=fr

outil de navigation (mots du corpus avec en rouge les supplémentaire, en bleu les variables étoilées); profils des types ; stat par classe (nombre de segment de texte par classe…) ; rapport.

Le logiciel sépare dans les graphiques de l'AFC les variables actives et les mots mais il est possible de récupérer le script R et de l'adapter.

Les fichiers générés sont sauvegardés dans un répertoire **: *nomcorpus*_alceste_1 :**

Le **corpus en couleur** permet de repérer à quel numéro de classe correspond le segment de texte classé.

Figure 12: Extrait du « corpus en couleur » issu d'une classification sur le corpus des réponses des étudiantes chinois (segment de texte ou textes ici spécifique des classes)

**** *n_241 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc1

clean fashionable healthy civilized

**** *n_242 *p_CHN *v_BJS *s_F *d_ART *e_0 *r_Inc3

developed economy beauteous environment linguistic diversity

**** *n_244 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc2

small area small population good environment beautiful scenery

---

**** *n_245 *p_CHN *v_BJS *s_M *d_ART *e_2 *r_Inc3

gleichschaltung contradiction civilized bright future

**** *n_246 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc2

free developed democratic lodgeable

**** *n_247 *p_CHN *v_BJS *s_M *d_ART *e_2 *r_Inc3

small cold foolish leisure expensive

**** *n_248 *p_CHN *v_BJS *s_M *d_ART *e_0 *r_Inc1

morden civilization classical architecture the homeland of white

EuroBroadMap, 2009

Pour chaque classe, on peut visualiser *le graphe de la classe* (cf. analyse de similitude) 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.

4) Analyse de Texte→ Analyse de similitude

Il s'agit d'une analyse des cooccurrences présentée sous formes de graphiques de mots associés (Analyse des Données Relationnelles). Les indices de similitudes proposés dans IRaMuTeQ sont ceux disponibles dans la librairie *proxy* de R (Meyer, Buchta).

Figure 13 : Paramétrage (par défaut) de l'analyse de similitude" dans IRaMuTeQ

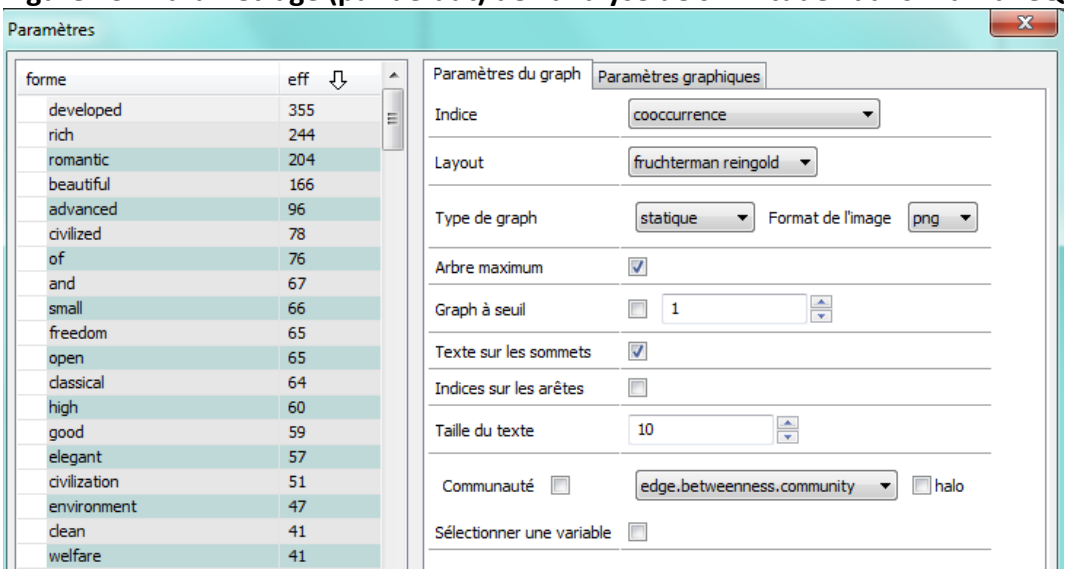
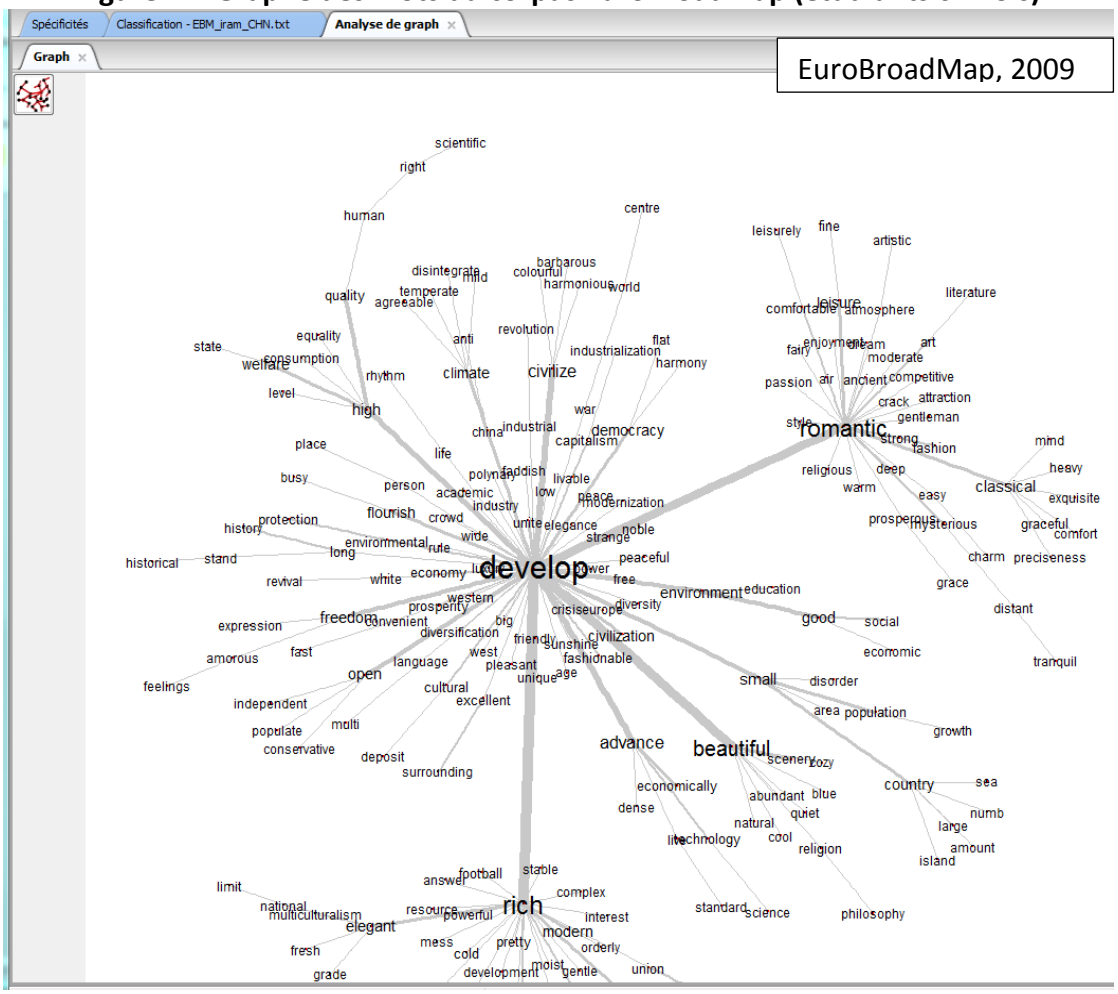


Figure 14: Graphe des mots du corpus EuroBroadMap (étudiants chinois)

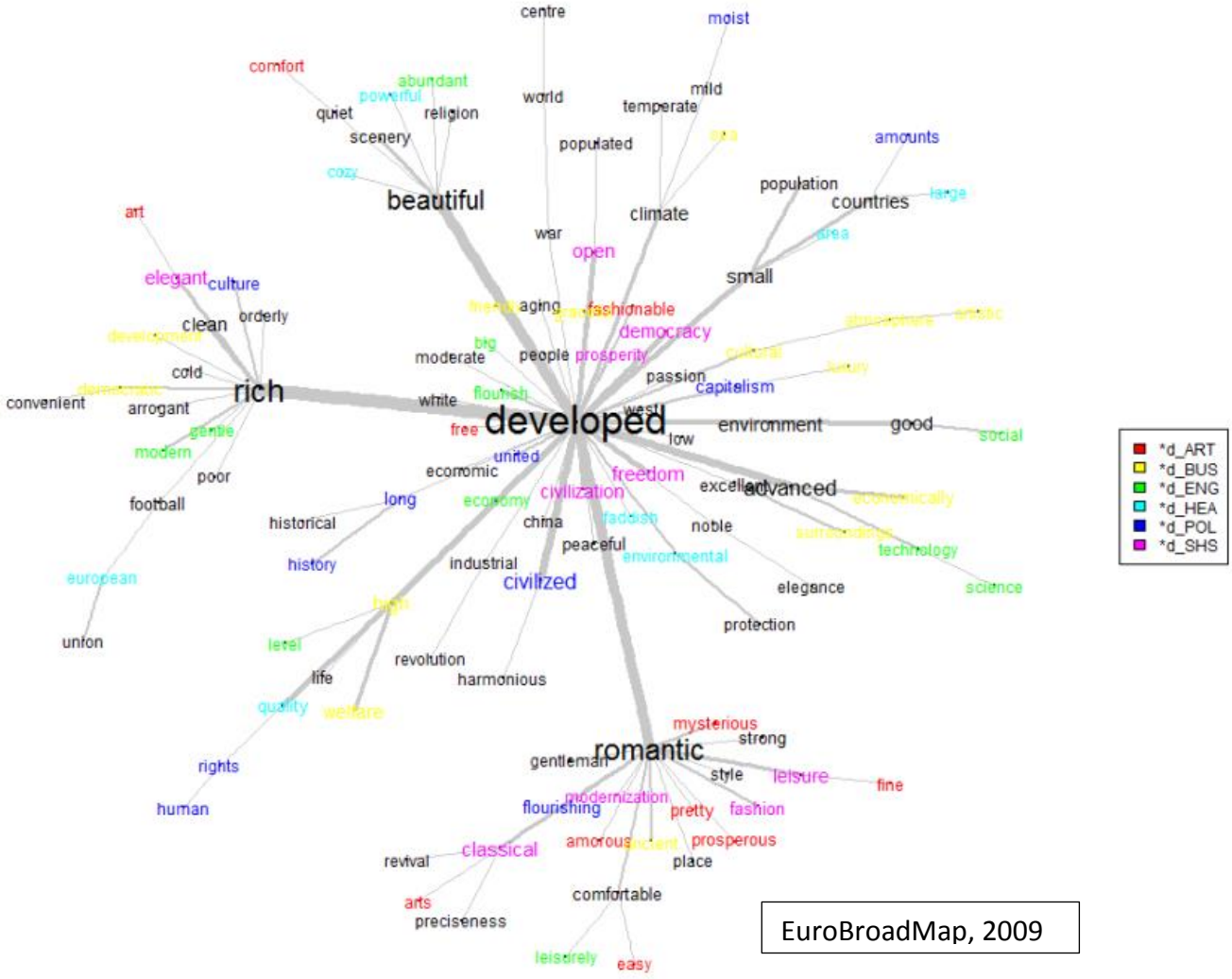


Possibilité d'exporter les graphes au format vectoriel (svg) ou pour gephi (format graphml) avec les coordonnées des points, la taille des sommets et leur couleur. (http://gephi.org)

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 (rose) sont spécifiques de la modalité (d_SHS).

Pour une meilleure visibilité, il est conseillé de cocher également « sélectionner les colonnes » qui permet de sélectionner les mots selon leur fréquence. Dans l'exemple ci-dessous, les mots ayant une fréquence supérieure à 6 ont été représentés.

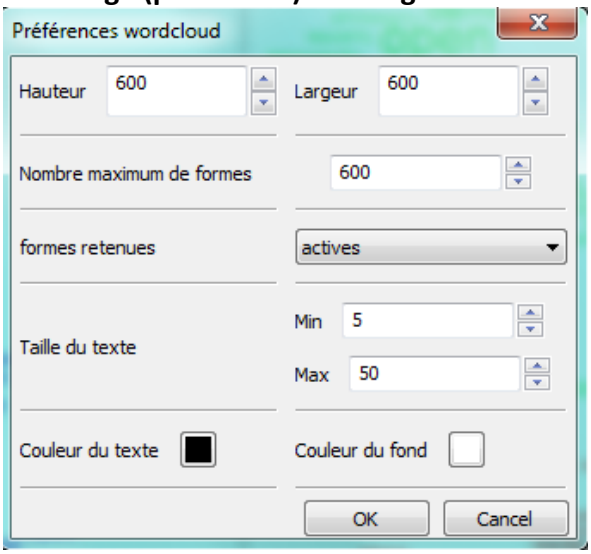
Figure 15: Graphe des mots du corpus EuroBroadMap (étudiants chinois) selon le domaine d'études



EuroBroadMap, 2009

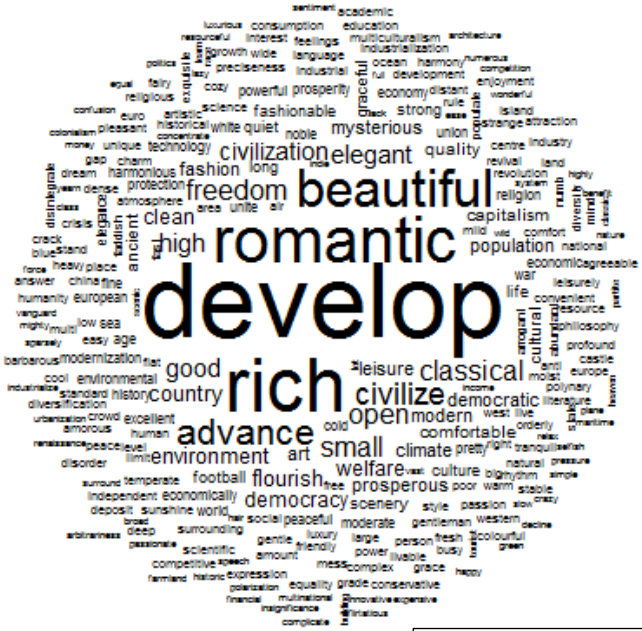
5) Analyse de Texte→ Nuage de mots

Figure 16: Paramétrage (par défaut) du nuage de mots dans IRaMuTeQ



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

Figure 17: Nuage de mots du corpus EuroBroadMap (étudiants chinois)



EuroBroadMap, 2009

Références

- http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf
- 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]