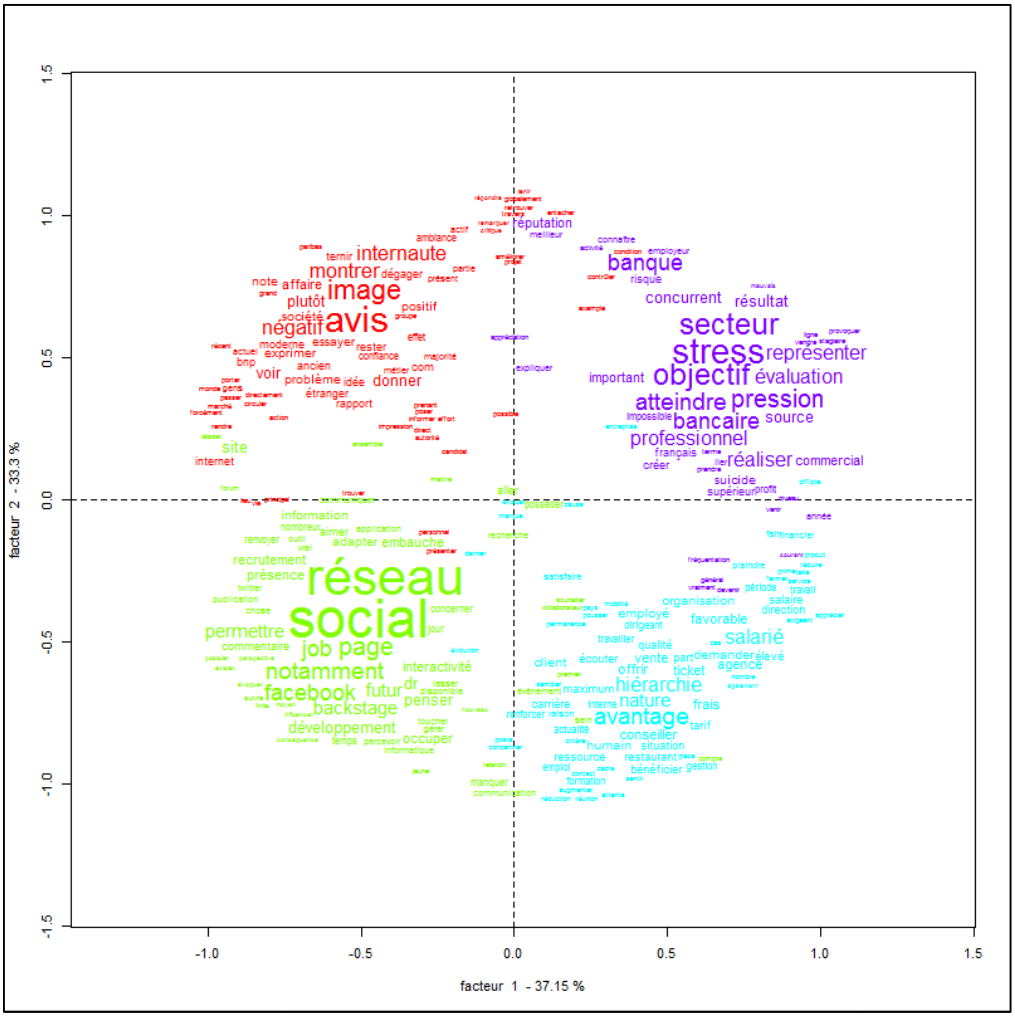
Initiation à la lexicométrie

Approche pédagogique à partir de l'étude d'un corpus

avec le logiciel

IRaMuTeQ



Iramuteq, logiciel libre développé par P. Ratinaud

Auteur du document : D. Pélissier

Date : mai 2016

Document associé : corpus sur les mots de proviseur des établissements de l'académie de Toulouse

Préambule

La relative facilité de l'usage de cet outil et la « beauté » des formes graphiques obtenues ne doivent pas faire oublier la nécessaire réflexion pour l'interprétation des données, réflexion qui, in fine, dépend de l'utilisateur.

Ce logiciel est libre (sources ouvertes), développé par P. Ratinaud et soutenu par P. Marchand. Sans leur investissement, largement bénévole, vous ne pourriez pas profiter de ce logiciel. Qu'ils soient chaleureusement remerciés pour cet énorme travail. De même, la gestion du forum dédié à ce logiciel par P. Ratinaud et L. Loubère est un élément essentiel dans la diffusion des usages de cet outil.

Merci aussi à P. Ratinaud et L. Wherlé pour leur relecture attentive ainsi qu'à Hélène Piment et aux étudiants du Dut Informatique de Rodez, promotion 2015-2017, pour leurs remarques pertinentes.

Iramuteq utilise le logiciel de traitement statistique R, qui est aussi libre.

Ce document a une vocation pédagogique. Il ne présente qu'une petite partie du logiciel. Il ne remplace évidemment pas la documentation complète d'Iramuteq rédigée par Lucie Loubère et Pierre Ratinaud (http://www.iramuteq.org/documentation).

Pour suivre ce document, vous devez avoir installé Iramuteq (et R, cf. documentation).

Les informations présentées engagent uniquement l'auteur du document.

Limites de ce document

La linéarité de la présentation masque les nombreux allers-retours d'une recherche menée avec cet outil et le lien entre corpus, résultats et méthode.

Certains choix (seuil de Chi^2, classification avant AFC) ne sont pas systématiques en lexicométrie.

Le corpus utilisé est plutôt de petite taille ce qui facilite l'approche pédagogique mais dissimule ainsi certains problèmes des gros corpus (création de sous-corpus notamment). C'est le « paradoxe pédagogique » (Lebart et Salem 1988 p. 51) de la démarche.

La démarche lexicométrique est volontairement présentée avec un outil particulier, Iramuteq (version 0.7 alpha 2), qui, malgré tous ces avantages, a ses propres limites.

Les méthodes utilisées ne couvrent qu'une infime partie des possibilités du logiciel ; d'autres fonctions permettent d'étudier d'autres formes de données (matrice et analyse prototypique etc.).

Enfin, la lexicométrie n'est pas la seule méthode d'analyse des discours. Elle n'empêche pas aussi de lire le corpus !

Table des matières

Définition et enjeux

Quelques éléments d'histoire

La lexicométrie peut être définie comme un : « *ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques sur le vocabulaire d'un corpus de textes* » Lebart et Salem 1988 p. 183.

Des mots proches de lexicométrie sont utilisés parfois comme : textométrie, logométrie etc. qui en étant proches sont sensiblement différents.

La lexicométrie s'est développée en France principalement depuis les années 70 avec notamment les travaux d'Etienne Brunet, Charles Muller, Pierre Guiraud ou encore Maurice Tournier[1] ou de statisticiens comme Jean-Paul Benzécri à l'origine de l'analyse factorielle de correspondance (Marchand 2013).

Mais l'essor de cette méthode est lié à celui des outils numériques et le développement des capacités de calcul des ordinateurs a favorisé la diffusion de la lexicométrie.

Deux auteurs ont contribué à l'explication et la formation : Ludovic Lebart et André Salem. Deux livres ont synthétisé les bases des méthodes utilisées : « Analyse statistique des données textuelles » en 1988 puis « Statistique textuelle » en 1994, ce dernier étant téléchargeable (voir webographie p. 31).

En parallèle, des logiciels implémentant ces traitements statistiques vont être développés comme Alceste[2] à partir de la fin des années 1980. Son concepteur Max Reinert, membre du CNRS, a ainsi permis l'automatisation des calculs ce qui a favorisé l'usage de la lexicométrie notamment dans les sciences humaines.

Plus récemment, des logiciels libres sont apparus comme TXM développé à partir de 2007 ou Iramuteq à partir de 2009.

Enjeux scientifiques et pédagogiques

La lexicométrie est utilisée pour les corpus volumineux de textes. Le développement d'internet et le phénomène du big data rendent d'autant plus pertinents l'usage de ce type d'outil (Barats, Leblanc et Fiala 2013). Mais l'importance de la taille d'un corpus étant subjective, la lexicométrie tend à être utilisée pour des corpus moins importants en complément ou remplacement de l'analyse de contenu par exemple. La lexicométrie est principalement utilisée pour une finalité scientifique mais est investie aussi par des consultants ou dans l'enseignement. Dans tous ces contextes, la rigueur du cadre statistique ne doit faire oublier la nécessaire qualité des analyses qui, in fine, détermineront la pertinence des travaux. De même, les résultats dépendent aussi de la qualité des données.

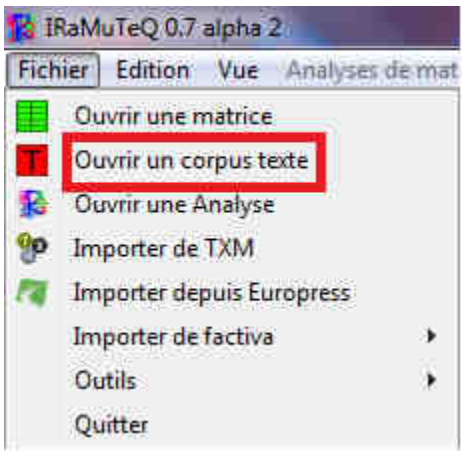
L'utilisation pédagogique relève d'une logique différente. La lexicométrie facilite l'analyse de discours divers (politiques, économiques (offres d'emploi, sites webs d'entreprise etc.), littéraires (extraits de romans, poèmes etc.)) dont le vocabulaire est lié à des représentations sociales (Ratinaud et Marchand 2015). Etre sensibilisé à ces variations conduit à mieux comprendre la complexité, la variété d'un discours, le sens des mots et de leurs relations au-delà de leurs fréquences.

Ce document en montrant à travers un exemple concret l'usage d'un logiciel de lexicométrie souhaite sensibiliser à cette méthode et à la complexité de la communication.

[1] Voir aussi : PINCEMIN Bénédicte, HEIDEN Serge (2008) - "Qu'est-ce que la textométrie ? Présentation", Site du projet Textométrie , http://textometrie.ens-lyon.fr/spip.php?rubrique80, consulté le 10 mars 2016.
[2] Site : http://www.image-zafar.com/Logiciel.html consulté le 10 mars 2016.

Construction et découverte du corpus

Commencer par ouvrir votre corpus[3] :



Organisation du corpus

Un corpus, dans cette situation, est un ensemble de textes.

Pour le traitement dans Iramuteq, l'ensemble des textes est regroupé dans un fichier texte (.txt) au format UTF8 et doit respecter un certain nombre de caractéristiques[4].

Exemple extrait : corpus de mots de proviseurs de sites internets de lycée

```
**** *nom_faure *type_lgt *dpt_ariege

L'objectif de ce site est de vous informer sur la vie pédagogique, éducative et culturelle de l'établissement mais aussi de vous donner l'envie d'aller à la rencontre de notre communauté scolaire.
```

Les textes commencent par 4 étoiles : ****

Une étoile désigne une variable ex. : *nom_faure

Le mot qui suit l'étoile est le nom de la variable ex. : nom

Pour les noms de variables et le contenu, et comme souvent en informatique, pas d'accents, d'espaces, de caractères spéciaux (+ - / * etc.).

A la suite du nom, est placé le contenu de la variable : ex. : faure

Ex. : *dpt_ariege désigne la variable dpt (département) qui correspond dans ce cas au département de l'Ariège.

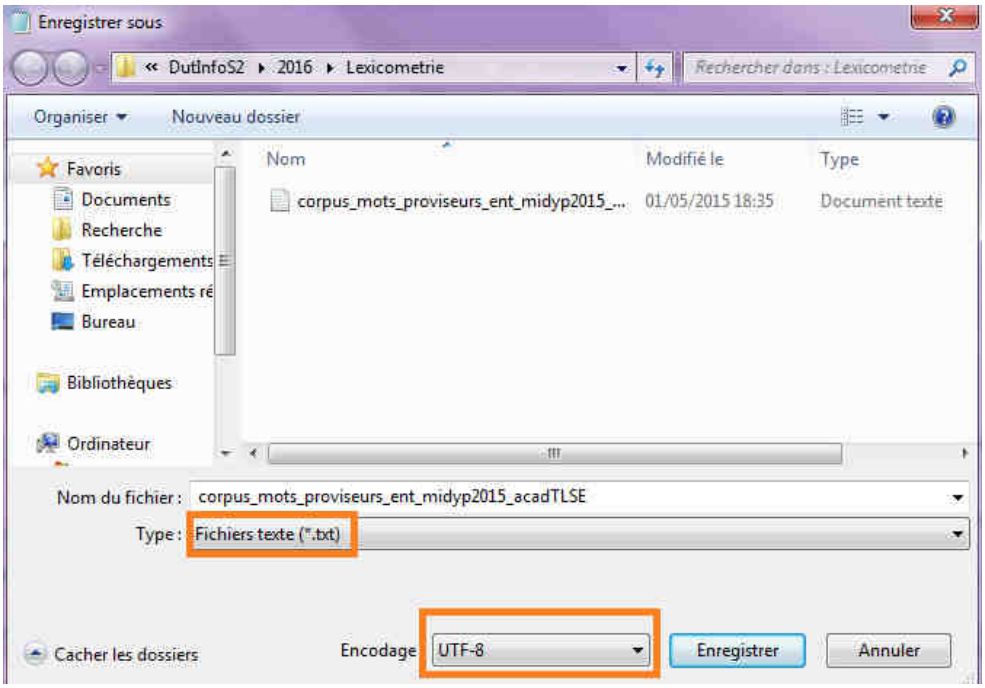
Ex. complet : **** *nom_faure *type_lgt *dpt_ariege

Texte spécifié par trois variables : la variable nom qui contient 'faure' pour ce texte, la variable type qui contient 'lgt' et la variable 'dpt' qui contient 'ariege'.

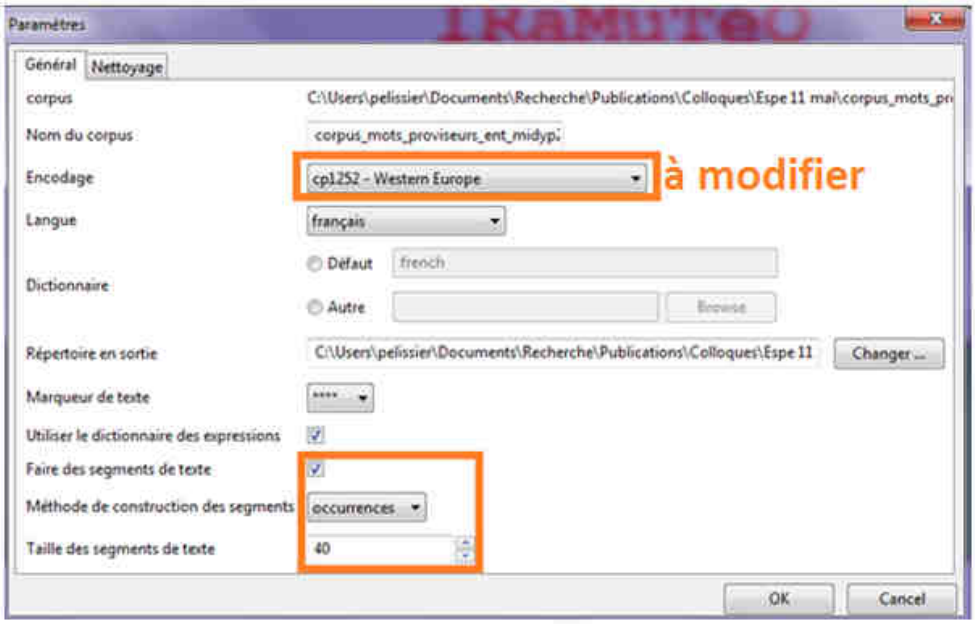
Ce texte est donc un mot de proviseur du lycée général et technique Faure du département de l'Ariège.

[3] Les pages écran correspondent à la version 0.7 alpha 2 d'Iramuteq.
[4] Il est recommandé de commencer le corpus par une ligne VIDE.

ATTENTION : les corpus[5] doivent être enregistrés au format UTF8 si vous utilisez le logiciel Bloc-notes[6] de Windows:



Iramuteq vous propose la fenêtre suivante :

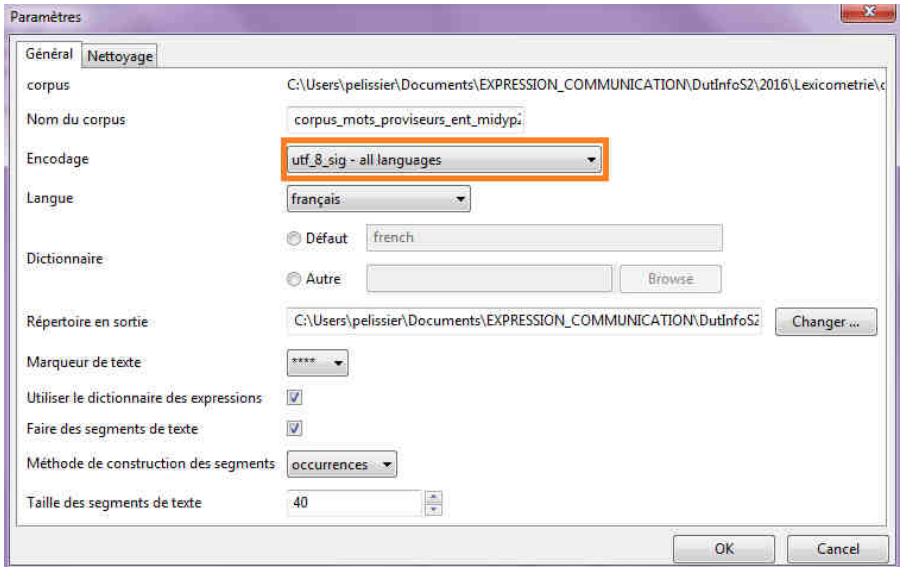


Les problèmes d'encodage se repèrent par un message d'erreur ou, plus sournois, par des remplacements de lettres en caractères spéciaux. Il faut alors surveiller les formes pour vérifier l'absence de problèmes.

[5] Il existe d'autres travaux de préparation du corpus : suppression des fautes d'orthographe, neutralisation de certaines formes (exemple : _IUT_ permet de ne pas analyser le mot « IUT »). Il est possible de modifier le dictionnaire : « les dictionnaires pour les différentes langues sont disponibles » documentation Iramuteq p.9
dans le répertoire .iramuteq/dictionnaires

[6] En effet, avec le Bloc-notes de Windows, les encodages disponibles sont limités (ANSI, unicode, unicode big endian et UTF-8). L'UTF-8 permet ensuite de faire correspondre l'encodage et les possibilités plus nombreuses d'Iramuteq. Libreoffice peut aussi être utilisé.

Vous devez modifier le format et sélectionner UTF8 pour obtenir la fenêtre suivante :



Les autres options ne sont pas modifiées. Les derniers éléments du formulaire permettent de segmenter le texte en occurrences tous les 40 mots[7]. Cette opération est très importante pour la suite des analyses. Le nettoyage permet de formater le corpus pour faciliter son analyse. Exemple : « l'entreprise » sera remplacé par « l entreprise » (suppression de l'apostrophe) ce qui permet d'analyser le mot entreprise. Enfin les dictionnaires permettent notamment de repérer certaines expressions et de lemmatiser (obtenir les racines des mots). La lemmatisation sera effectuée lors du traitement statistique du corpus. Si vous avez bien formaté votre corpus (voir ci-dessus), vous obtenez la fenêtre suivante :

Description du corpus	
Nom	corpus_mots_proviseurs_ent_midyp2015_acadTLSE_corpus_2
Langue	french
Encodage	utf_8_sig
originalpath	C:\Users\pelissier\Documents\EXPRESSION_COMMUNICATIO
pathout	C:\Users\pelissier\Documents\EXPRESSION_COMMUNICATIO
date	Thu Dec 17 10:59:18 2015
time	0h 0m 1s
Paramètres	
ucemethod	1
ucesize	40
keep_caract	^a-zA-Z0-9àÀâÂäÄáÁåÅãéÉèÈêÊëËìÌîÎïÍóÓòÒôÔöÖõÕøØùÙû
expressions	1
Statistiques	
Nombre de textes	57
Nombre de segments de texte	403
occurrences	14162
Nombre de formes	2715
Nombre d'hapax	1483 - 54.62 % des formes - 10.47 % des occurrences

Dans ce cas, le corpus est formé de 57 textes, 403 segments de textes et 14162 occurrences (mots dans ce cas).

[7] Pourquoi pas 42 mots ? La priorité est donnée à la ponctuation ce qui explique des longueurs moyennes de segment inférieures à 40. Le seuil de 40 correspond à une pratique. Pour comparer des traitements de corpus, il est préférable de choisir un même seuil.

Iramuteq a segmenté le texte tous les 40 mots. Le nombre de segments devrait être de 354 (14162/40) ; or il est, dans ce cas, de 403. En effet, Iramuteq découpe le corpus en respectant la ponctuation : extrait documentation Iramuteq (L. Loubère, P. Ratinaud, http://www.iramuteq.org/documentation)

« *Les segments de texte sont construits à partir d'un critère de taille et de ponctuation. Iramuteq cherche le meilleur ratio taille/ponctuation (par ordre de priorité, les ".", "?" "!" en premier, puis en second ";" et les ":" en troisième la virgule et en dernier l'espace). L'objectif est d'avoir des segments de tailles homogènes en respectant le plus possible la structure du langage*. »

Le nombre de formes correspond aux formes de mots trouvés différentes. Il est ainsi inférieur au nombre d'occurrences.

Ex. : « Le lycée est un lieu de vie. Lycée et élèves sont liés de fait».

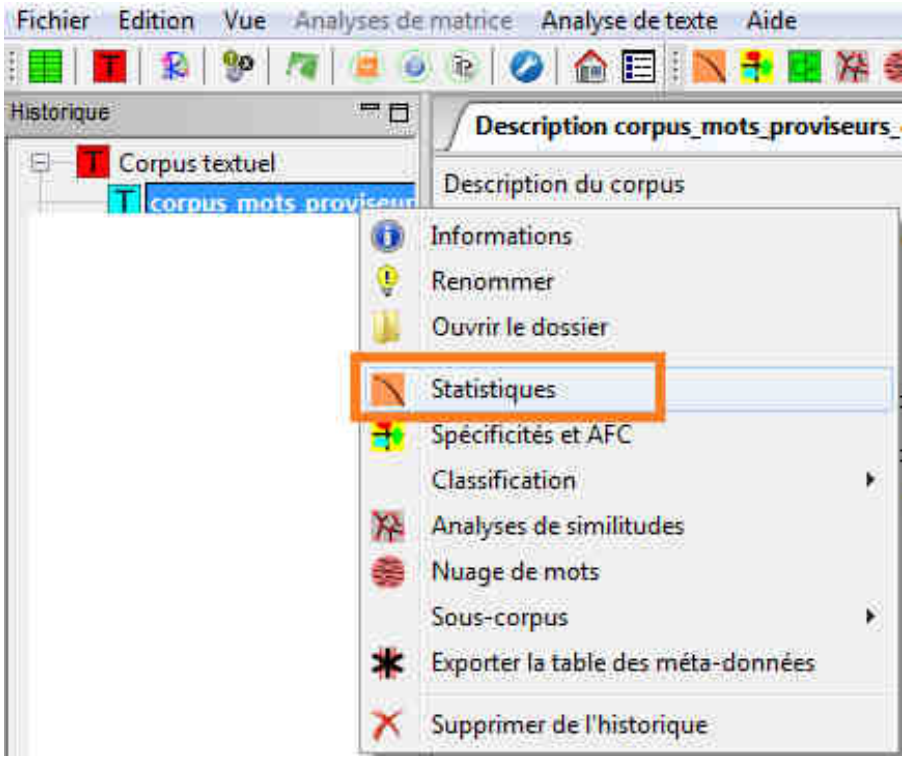
Contient 14 occurrences et 12 formes ('lycée' et 'de' sont répétés).

Les hapax sont des occurrences dont la fréquence est unique. Ex. si le mot « hexadécimal » apparait une seule fois dans un corpus, alors, pour ce corpus, « hexadécimal » est un hapax.

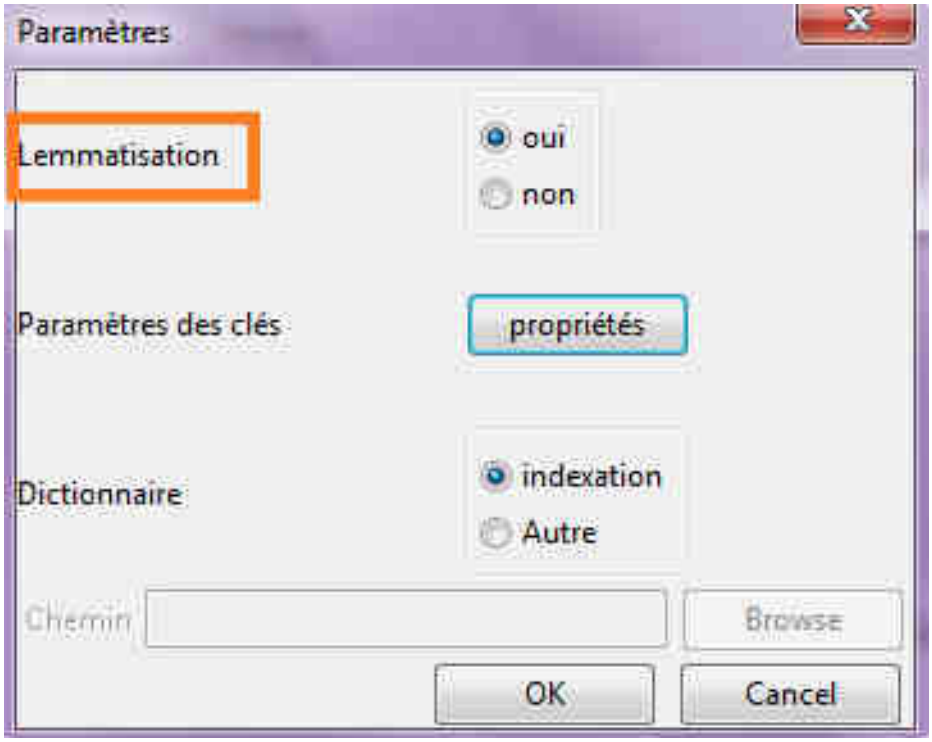
Dans l'exemple, le corpus contient 1483 occurrences dont la fréquence est de 1 ce qui représente 1483/14162 du corpus soit 10,47% des occurrences et 1483/2715 soit 54, 62% des formes.

Description du corpus

Pour obtenir, les premières statistiques sur le corpus, sélectionnez la ligne « Statistiques » :



Le logiciel vous propose la fenêtre suivante :



Les options sont laissées par défaut dans cet exemple.

LEMMATISATION : les occurrences sont réduites à leur racine ; les verbes sont ramenés à l'infinitif, les noms au singulier et les adjectifs au masculin singulier.

Ex. : « le lycée organise des portes ouvertes » donnera les lemmes suivants : le, lycée, organiser, de, porte, ouvert.

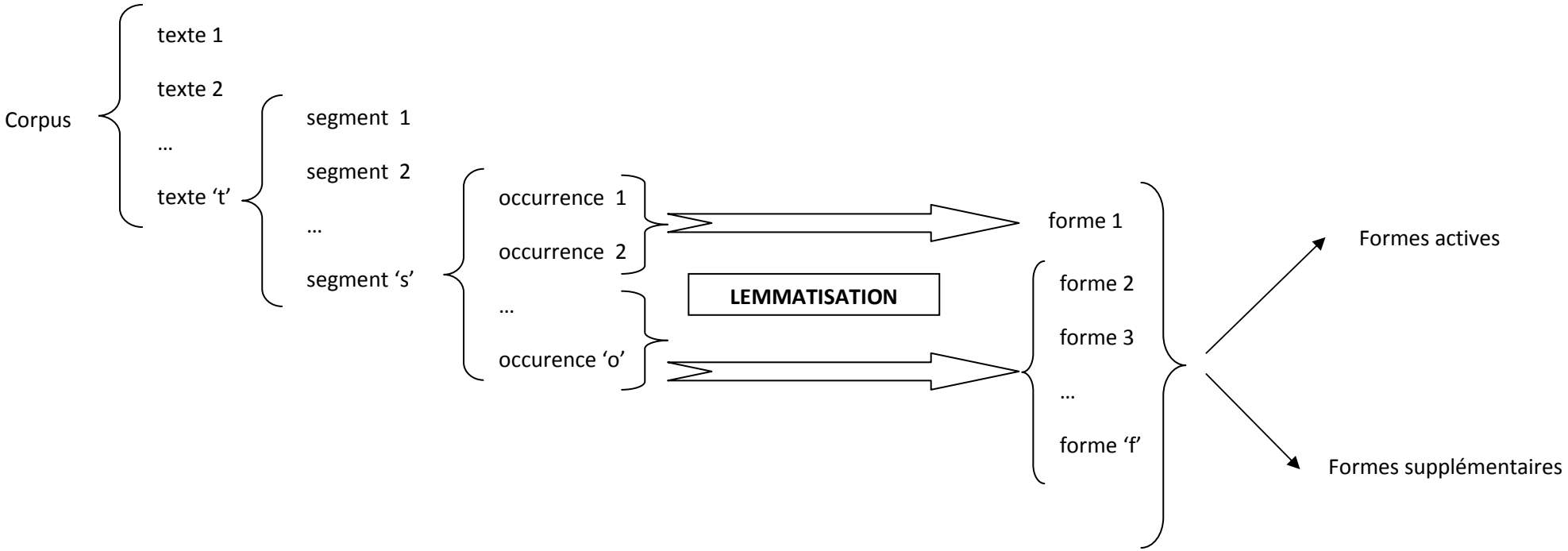
Remarque : les clés permettent de spécifier les formes qui seront analysées. - ce qui est mis en actif par défaut (codé 1) : adjectifs, adverbes, formes non reconnues, noms communs et verbes. - ce qui est mis en supplémentaire par défaut (codé 2) : mots outils. (source : Elodie Baril et Bénédicte Garnier, http://www.iramuteq.org/documentation/fichiers/Pas%20a%20Pas%20IRAMUTEQ_0.7alpha2.pdf)
Le logiciel traitera toutes les formes mais il distinguera les formes actives (clé=1) les formes supplémentaires (clé=2).

Vous pouvez aussi observer les mots du dictionnaire classés par catégorie : verbe, adjectif, adverbe etc.

Le dictionnaire sera, dans ce document, celui utilisé par défaut par Iramuteq. Il peut être modifié (documentation Iramuteq p. 9).

Les paramètres des clefs permettent de choisir, par catégories grammaticales, quelles formes seront considérées comme actives ou supplémentaires. Par exemple, dans une classification de Reinert, seules les formes actives seront utilisées. Si, dans votre corpus, les onomatopées sont importantes, vous devrez modifier les clés.

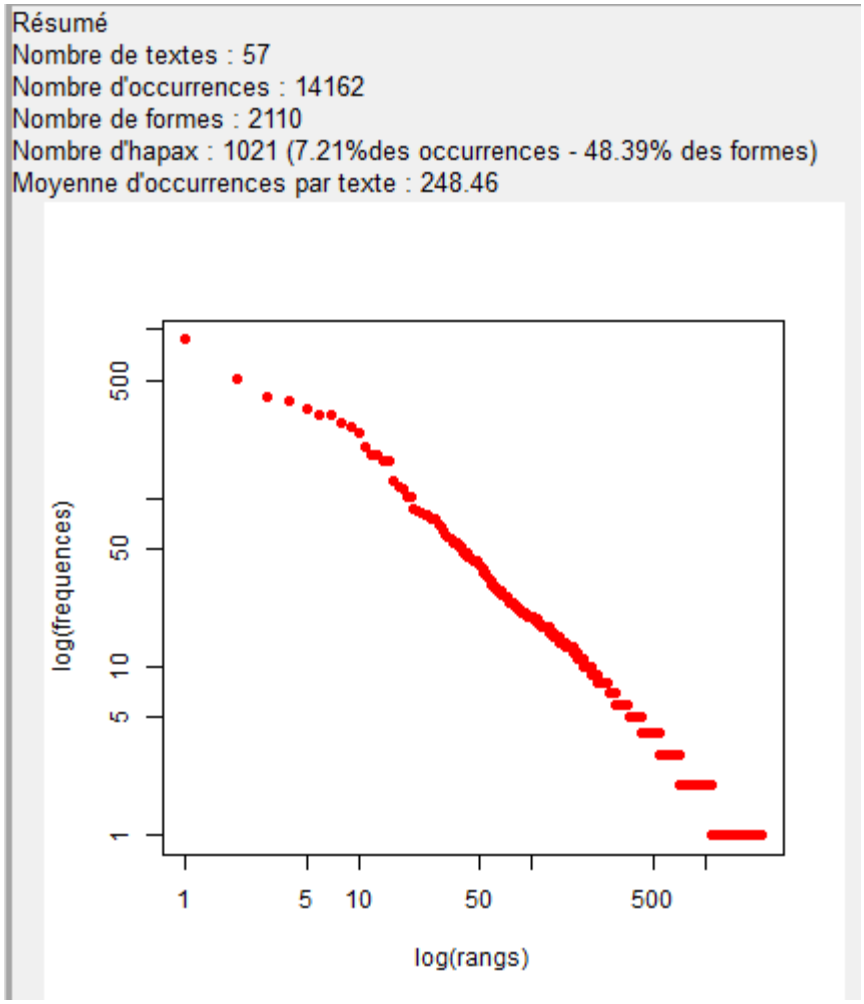
Synthèse du vocabulaire



« **Segment** : toute suite d'occurrences consécutives dans le corpus et non séparés par un séparateur de séquence.

Lemmatisation : regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante : les formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier, les formes élidées à la forme sans élision. » Lebart et Salem 1988

Vous obtenez notamment le résultat suivant :



Vous pouvez vérifier que le nombre de textes est identique à l'étape précédente (indexation) comme le nombre d'occurrences. Par contre le nombre de formes est inférieur en raison de la lemmatisation et du dictionnaire utilisé. Le nombre d'Hapax est aussi légèrement inférieur.

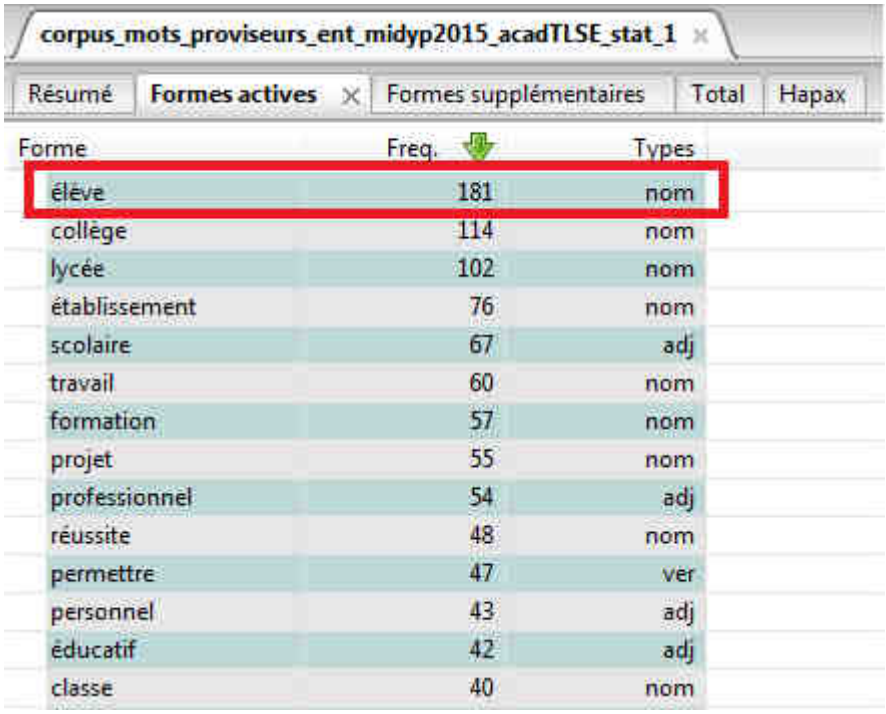
Le graphique présente en abscisse les logarithmes des rangs et en ordonnées les logarithmes des fréquences des formes (L. Loubère et P. Ratinaud, documentation). Les valeurs indiquées sont par contre exprimées dans leur unité de départ. Les mots les plus fréquents sont ceux qui ont le rang le plus élevé. Par exemple, le mot le plus fréquent d'un corpus aura le rang 1. Ex. : le mot « de » est très fréquent (500) et a le rang 1.

L'échelle logarithmique permet de réduire l'échelle pour les fréquences et rangs élevés. La courbe est ainsi souvent une droite décroissante[8]. Les hapax (fréquence 1) ont des rangs identiques (ex aequo) d'où le trait horizontal : même fréquence (1) et plusieurs rangs ex-aequo.

Des cas particuliers peuvent être intéressants à observer si la courbe est stable au départ et décroissante ensuite : certaines occurrences sont majoritaires.

[8] Elle illustre aussi la « loi » de Zipf qui stipule que le produite rang*fréquence est une constante (Lebart et Salem 1988, p. 34)

Des onglets permettent de repérer certaines formes utilisées fréquemment :

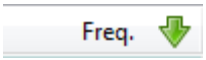

corpus_mots_proviseurs_ent_midyp2015_acadTLSE_stat_1

Résumé | Formes actives | Formes supplémentaires | Total | Hapax

Forme	Freq.	Types
élève	181	nom
collège	114	nom
lycée	102	nom
établissement	76	nom
scolaire	67	adj
travail	60	nom
formation	57	nom
projet	55	nom
professionnel	54	adj
réussite	48	nom
permettre	47	ver
personnel	43	adj
éducatif	42	adj
classe	40	nom

La forme active la plus utilisée est, pour ce corpus, le mot « élève » utilisé 181 fois, nettement devant collège ou lycée. La dernière colonne fournit le type du mot (nom dans ce cas).

Les formes supplémentaires présentent des mots de liaison (de, et, en etc.) ou des déterminants (la, le etc.). Enfin les happax sont les formes dont la fréquence est de 1.

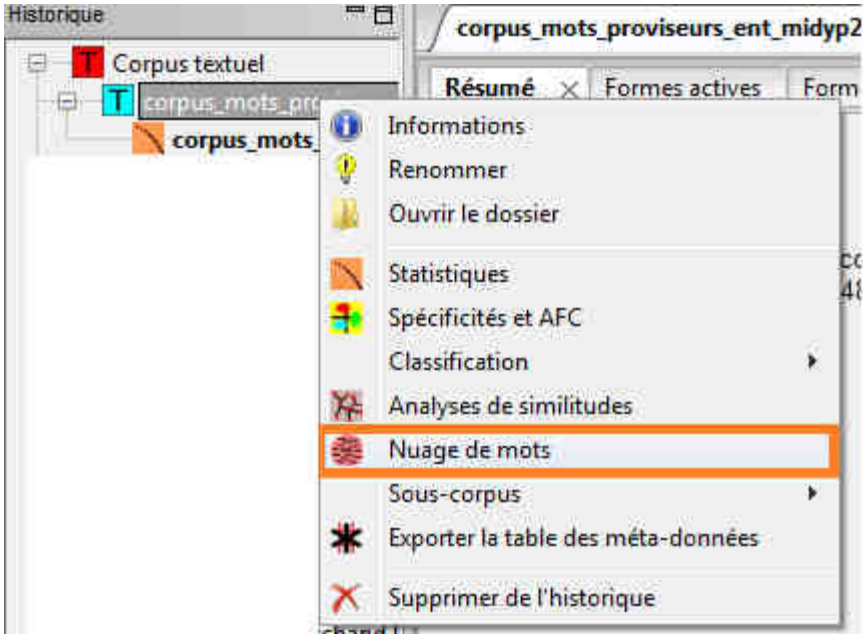
La flèche sur la fréquence : 

permet de classer par ordre croissant ou décroissant.

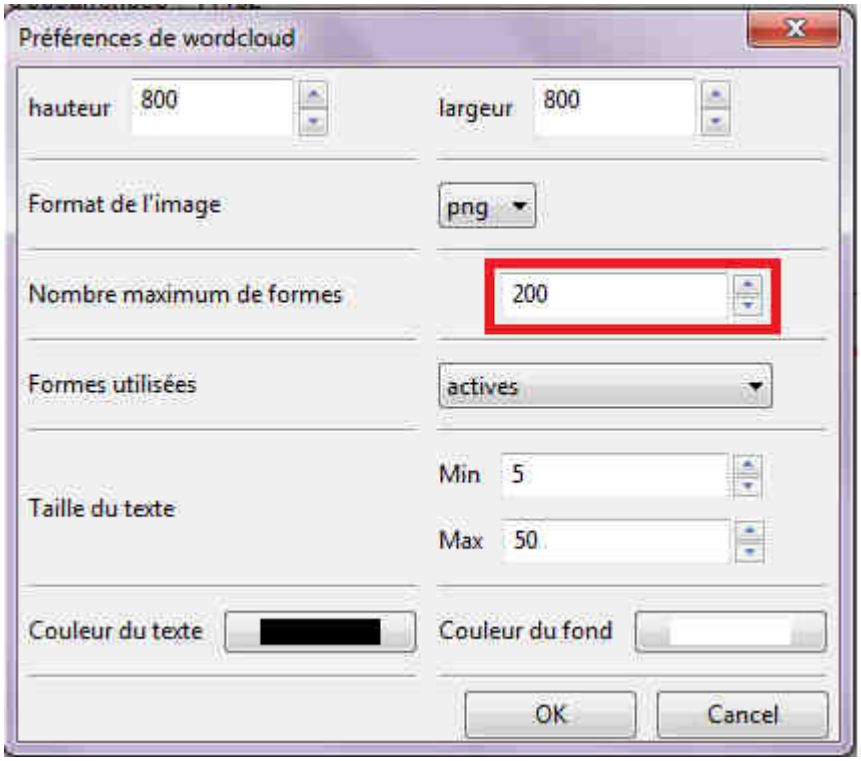
ANALYSE DU CORPUS

NUAGE DE MOTS

Nous commencerons par réaliser un nuage de mots :



Après avoir, comme pour les statistiques, validé les options par défaut de lemmatisation, la boite de dialogue suivante apparait :



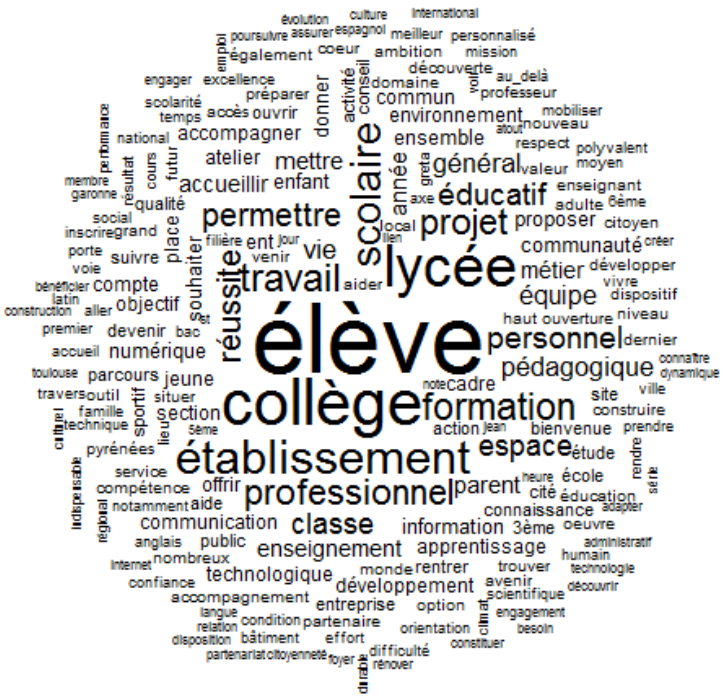
Le nombre de mots a été abaissé à 200 pour faciliter la lecture.

Les autres options sont laissées par défaut et notamment :

- les formes utilisées sont actives (pour ce corpus, les formes supplémentaires ne sont pas étudiées mais ce n'est pas une règle générale de lexicométrie) ;
- les couleurs sont contrastées et permettent aussi une impression en noir et blanc.

La fenêtre suivante permet de visualiser les fréquences des 200 premières formes. Comme certaines sont très peu fréquentes (corpus relativement petit), on peut sélectionner certaines formes plus fréquentes.

Le résultat est le suivant :



Ce schéma, très répandu sur internet, est un outil de lexicométrie qui représente simplement les fréquences des formes en les corrélant à leur taille. Elles sont ainsi liées à la description statistique du corpus montrée précédemment : les mots élèves, lycée, collège, scolaire, travail etc. sont les plus fréquents.

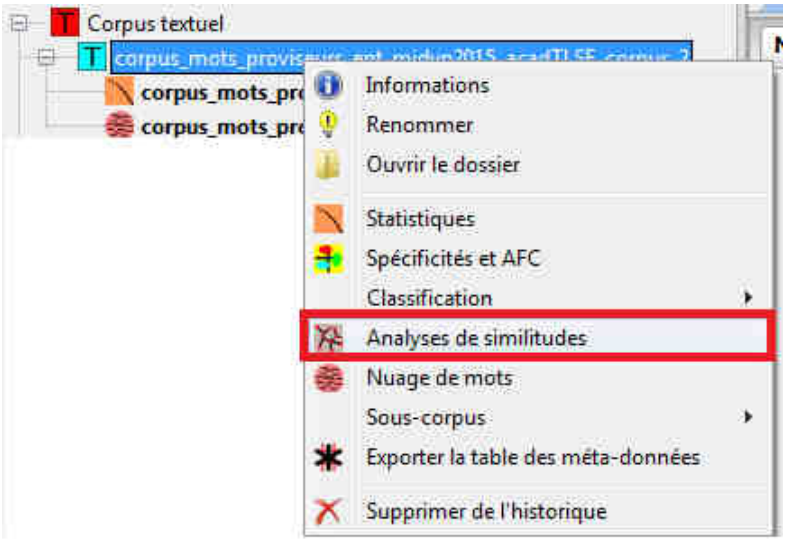
Mais cet outil ne montre pas les liens entre les formes ce qui est pourtant tout l'intérêt de la segmentation du texte.

Une autre technique permet cette étude : l'analyse de similitude.

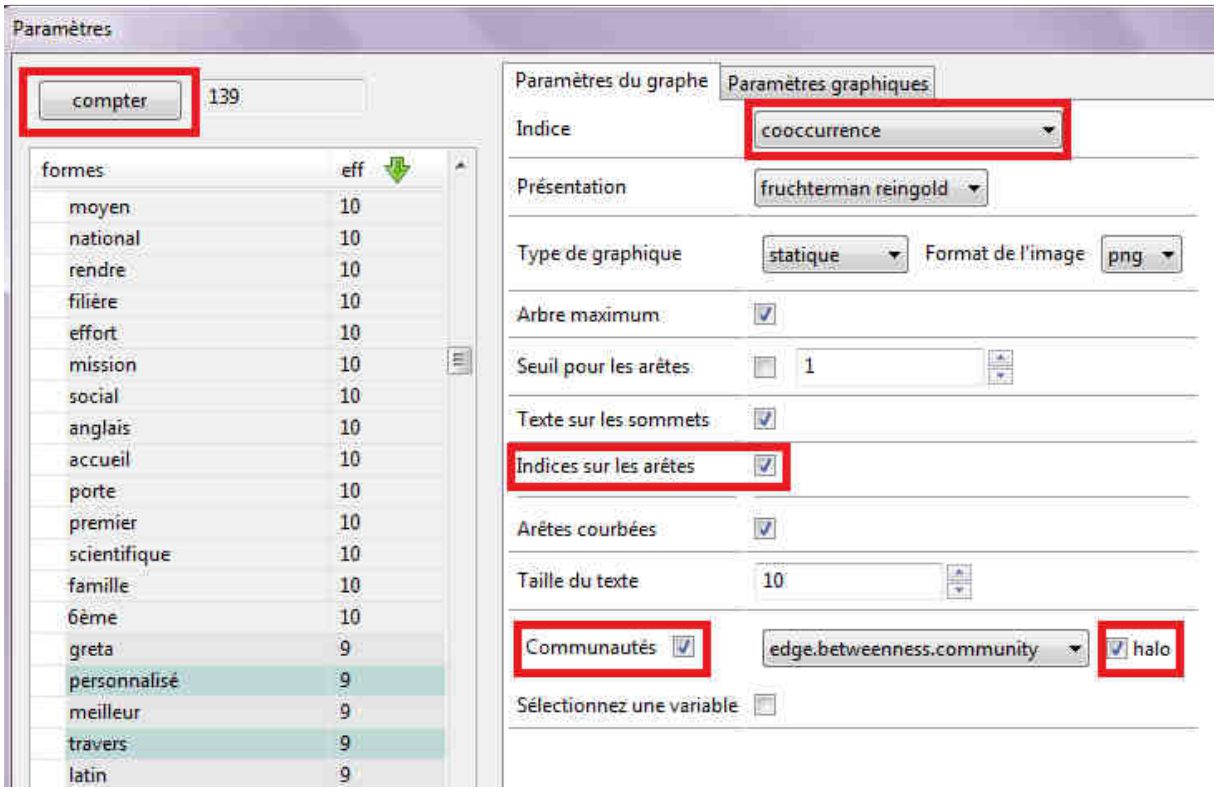
A noter : toutes les opérations réalisées sur le corpus créent une ligne de menu qui apparait en dessous de votre corpus ; jusqu'ici, nous obtenons :

ANALYSE DE SIMILITUDE

Pour réaliser une analyse de similitude, sélectionnez la ligne suivante :



La boite de paramètre est modifiée de la façon suivante :



L'indice utilisé est celui de la cooccurrence[9] qui peut être définie de façon générale comme : « apparition simultanée de deux ou plusieurs éléments ou classes d'éléments dans le même discours » (CNRTL : http://www.cnrtl.fr/definition/cooccurrence). Pour Iramuteq, sauf cas particulier, la cooccurrence est déterminée au niveau du segment de texte. Ainsi, l'indice de cooccurrence correspond au comptage du nombre de segments dans lesquels une forme est associée à une autre. Par exemple, un indice de cooccurrence de valeur « 10 » entre deux formes signifie que ces deux formes apparaissent ensemble dans 10 segments de texte.

[9] Pour en savoir plus : http://theses.ulaval.ca/archimede/fichiers/22376/ch05.html Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés Réhel, Simon. Maîtrise Maître ès sciences (M.Sc.).

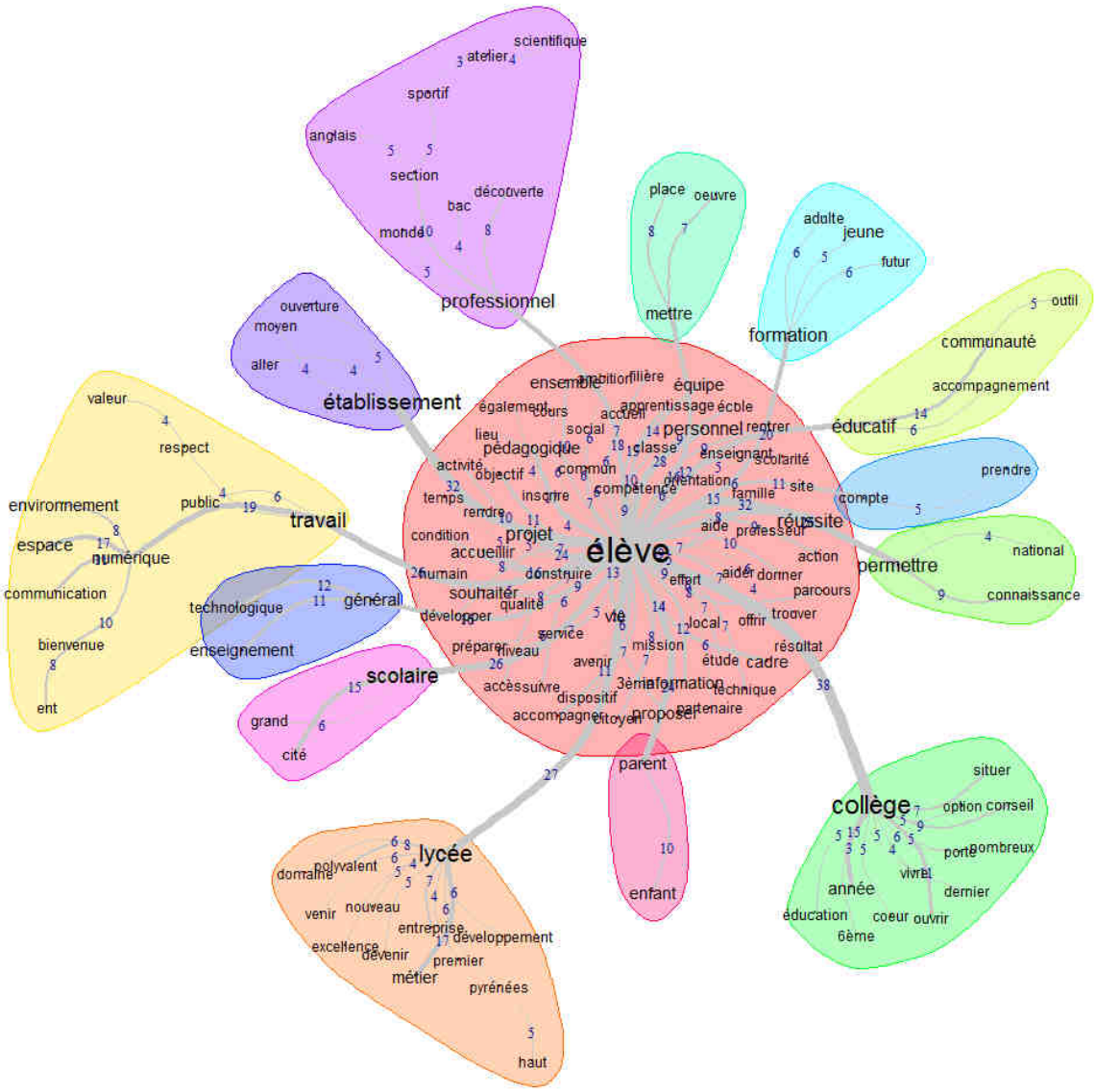
Cet indice sera visible sur les arêtes (les éléments graphiques reliant les formes) en cochant l'option adaptée.

En sélectionnant, « Communautés » et « halo », l'ADS (analyse de similitude) est colorée ce qui facilite son interprétation.

Pour ce corpus, les formes sélectionnées sont celles ayant une fréquence supérieure à 10[10] soit 139 formes (sélectionner les formes tout en cliquant sur la touche MAJ puis cliquer sur le bouton « compter » pour obtenir cette dernière information).

PS : il est possible de sélectionner une variable pour faire apparaître le lien entre la forme et la variable (test du Chi²[11]).

On obtient alors le graphique suivant :



Analyse de similitude, fréquence >=10, arbre maximum, indice de cooccurrence

[10] Ceci n'est pas une règle générale de lexicométrie. Il est possible d'analyser l'ensemble des formes mais l'ADS sera alors plus difficile à lire.

[11] Test statistique permettant d'évaluer l'indépendance de deux variables : http://grasland.script.univ-paris-diderot.fr/STAT98/stat98_8/stat98_8.htm. Voir aussi en annexe pour une application adaptée au cas présenté.

La taille des traits est proportionnelle à l'indice de cooccurrence des formes : par exemple, le lien entre élève et collège (38) est plus fort qu'entre élève et lycée (27). D'autres liens paraissent structurer le discours : scolaire (26), éducatif (20), parent (24) et de nombreuses relations[12] ont un faible indice de cooccurrence (autour de 5)[13].

La principale communauté de mots est structurée autour du mot « élève ». Trois autres se détachent : collège, lycée et numérique.

Les icônes suivantes :

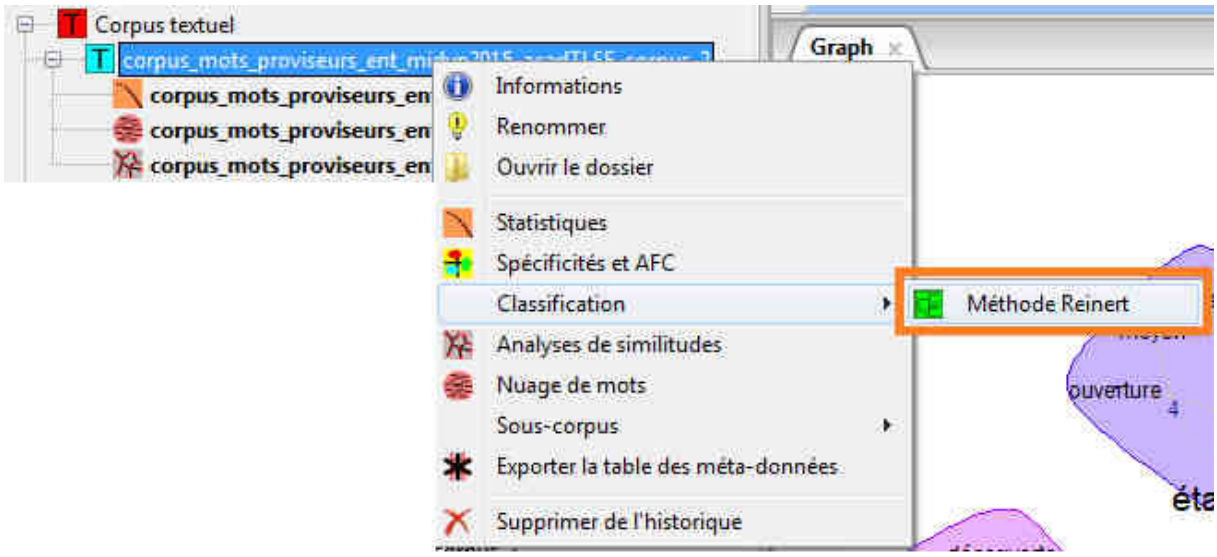
permettent de revenir aux paramètres de l'ADS (icône supérieure) et l'icône export permet de récupérer le fichier image. Si la représentation de l'AS est difficile à lire, vous pouvez aussi jouer sur l'option « *Seuil pour les arêtes* » qui permet de spécifier le seuil minimum d'affiche des cooccurrences.

A noter : tous les fichiers images sont stockés dans un sous répertoire créé automatiquement par Iramuteq à l'emplacement de votre fichier texte[14].

L'ADS a fourni certaines informations mais de nombreuses données sont encore trop segmentées. Un autre outil permettra de mieux lire ces informations.

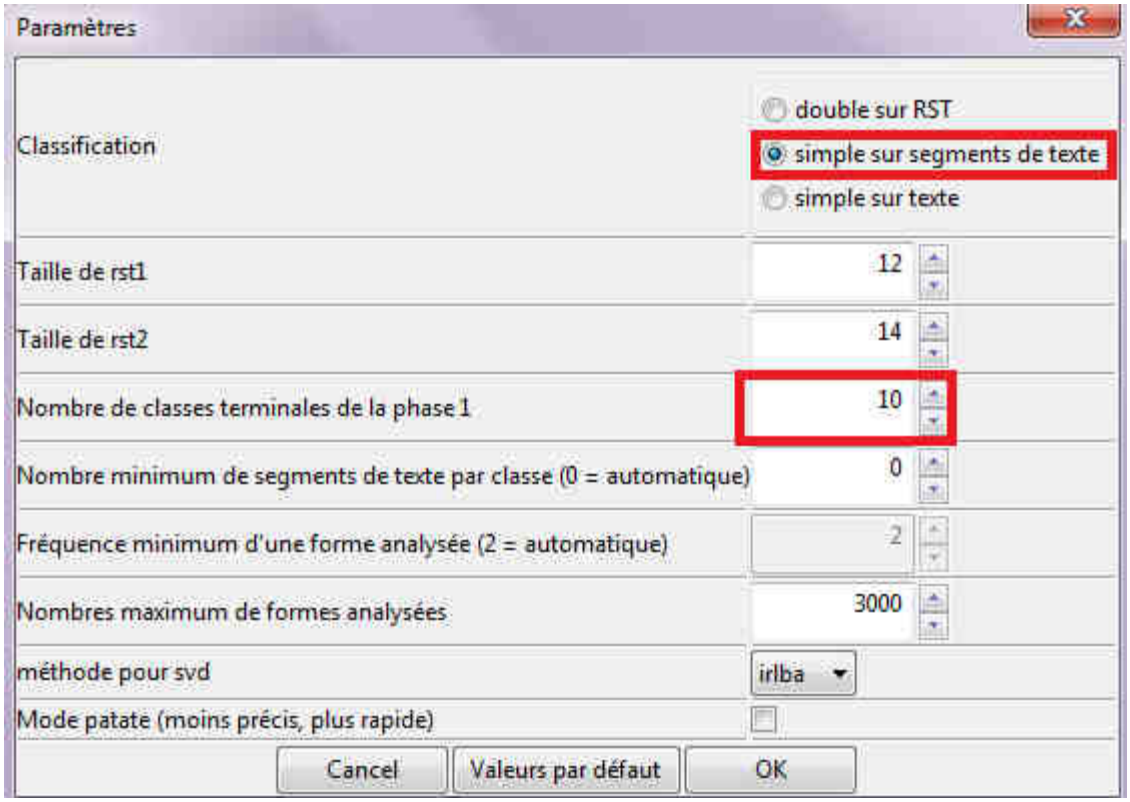
CLASSIFICATION DE REINERT

Pour réaliser cette classification, sélectionnez le menu suivant :



[12] Il peut exister aussi des relations non repérées entre des mots de communautés différentes : lycée et travail par exemple.
[13] Les cooccurrences sont parfois difficilement lisibles sur le graphique. Les données précises sont stockées dans le fichier « graph_1.graphml » dans le répertoire de l'analyse de similitudes.
[14] Il est possible d'afficher la matrice de l'analyse de similitude : fichier Rdata.rdata dans le répertoire, ouvrable avec le logiciel Rstudio https://www.rstudio.com/products/rstudio/download/

Dans la boite des paramètres :



Attention : le nombre minimum de segments de texte par classe par défaut n'est pas de zéro car ce nombre, comme spécifié, est utilisé pour réaliser automatiquement la classification (« par défaut ce nombre est égal au nombre de segments de texte divisé par le nombre de classes terminales pour la classification simple » documentation Iramuteq p. 21). Si vous souhaitez obtenir l'ensemble des classes possibles, vous devez entrer un seuil de 1. Il est alors possible ensuite de choisir un nombre minimum mieux adapté à votre corpus.

La classification de Reinert permet de classer les formes dans des classes de formes regroupées selon leur indépendance mesurée par un test au Chi².

Exemple de tableau lexical :

	Segment 1	Segment 2	…	Segment n
De	1	0		1
Lycée	1	1		1
Collège	0	1		0

1= la forme est présente au moins une fois dans un segment

0= la forme n'est pas présente dans le segment

Les paramètres taille rst1 (rst1= « Regroupement de Segments de Texte ») et taille rst2 (12 et 14 par défaut) ne sont utilisés que dans une classification double et correspondent au nombre minimum de formes pleines par ligne du tableau. Les segments qui auront le même profil (dans le tableau précédent, suite de 0 et 1 proche) seront regroupés en classe. Iramuteq analyse bien des segments (cotextes) et pas des mots isolés[15]. On obtient alors un dendogramme, arbre des classes. Le nombre maximum de classes créées peut être augmenté pour les corpus importants. De même, le mode « patate » est adapté aux gros corpus.

[15] Dans cette approche de la lexicométrie, on analyse des **textes** découpés en **cotextes** replacés dans un **contexte** de production.

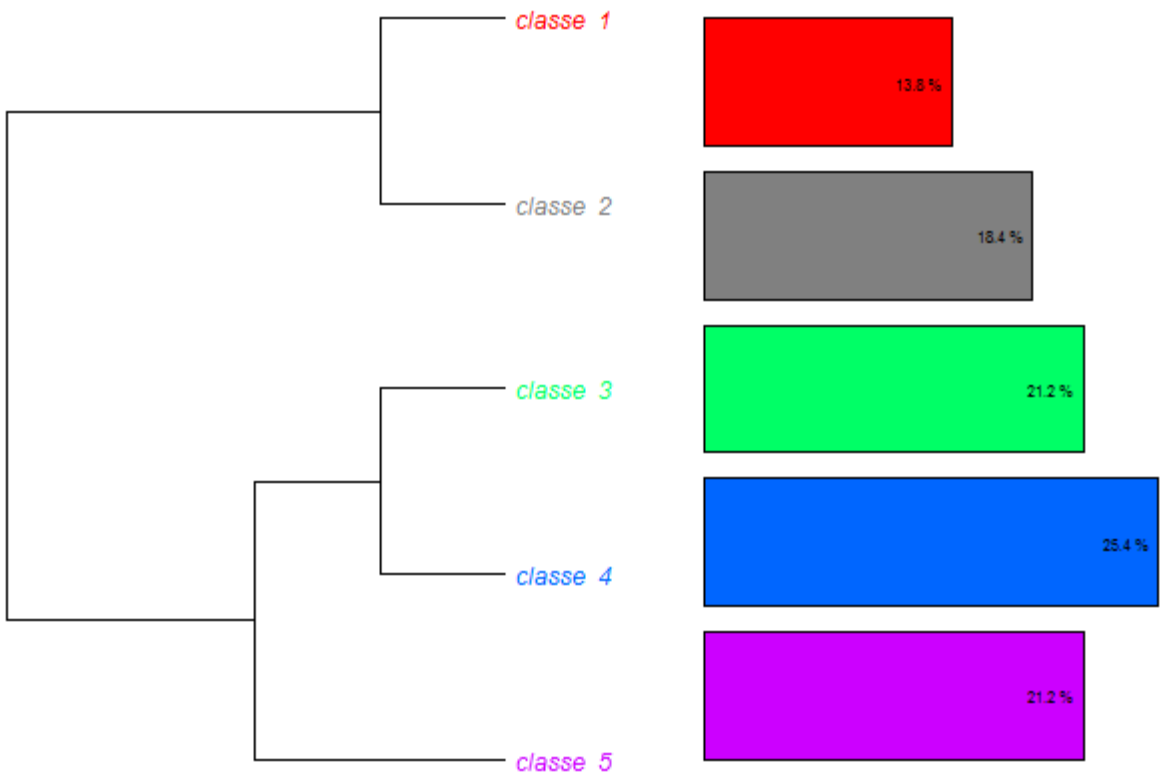
Le résultat est le suivant :

```
Nombre de textes: 57
Nombre de segments de texte: 403
Nombre de formes: 2715
Nombre d'occurrences: 14162
Nombre de lemmes: 2110
Nombre de formes actives: 1803
Nombre de formes supplémentaires: 307
Nombre de formes actives avec une fréquence >= 3: 570
Moyenne de formes par segment: 35.141439
Nombre de classes: 5
354 segments classés sur 403 (87.84%)

############################
temps : 0h 0m 15s
############################
```

Le corpus étant plutôt petit, le temps de traitement (avec cet ordinateur) est plutôt rapide : 15 s. On retrouve les statistiques précédentes sur le nombre de textes, segments etc. L'algorithme a pu déterminer 5 classes et 87, 84 % des segments ont pu être classés ce qui est démontre une bonne qualité de l'analyse[16].

Les classes sont les suivantes :

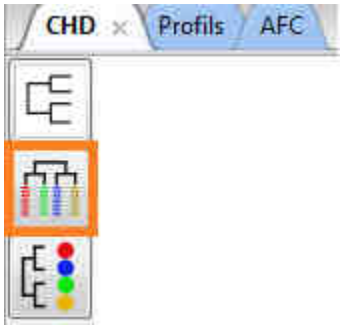


Deux sous-ensembles de classes se distinguent : classes 1 et 2 d'une part et classes 3, 4 et 5 d'autre part. **Il est possible que les classes n'apparaissent pas exactement dans cet ordre et avec ces couleurs selon votre environnement numérique. Il faudra alors légèrement adapter la suite du TD en fonction de vos données.** Ont été déterminées les classes 1 et 2 puis un autre ensemble qui a été découpé en classe 5 puis deux dernières classes 3 et

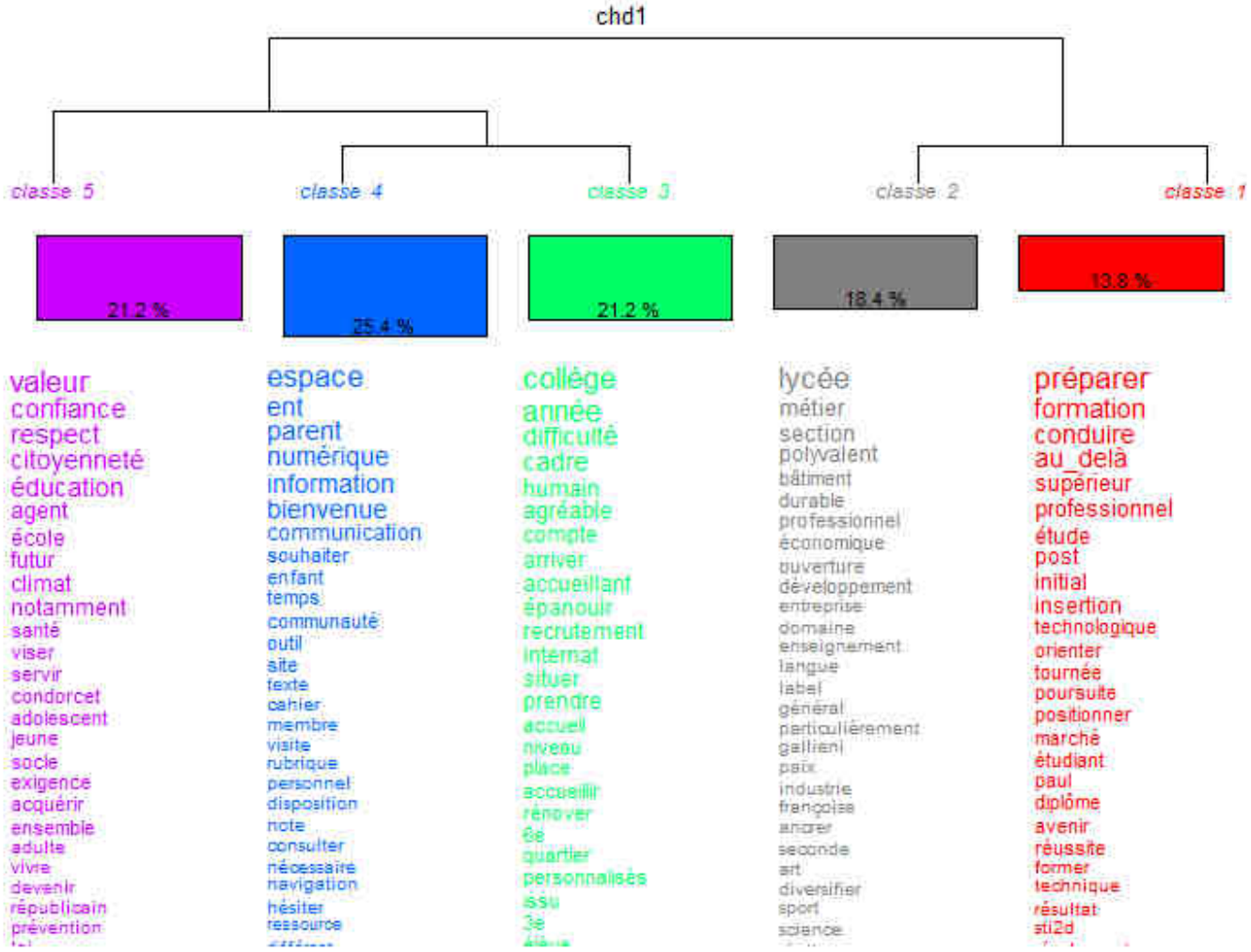
[16] Une valeur inférieure à 60% pour les discours écrits peut indiquer une analyse trop limitée. Les discours oraux (moins homogènes) peuvent cependant être analysés avec un taux de classement autour de 60%, seuil issu de la pratique.

4[17]. Les pourcentages représentent la quantité d'information résumée pour chaque classe. La classe 5 regroupe 21,2% des données (sans les classes non retenues). La somme de ces pourcentages est ainsi égale à 100%.

Les icones suivantes :



Permettent de présenter différemment les classes sous forme d'arbres[18] (dendogrammes) :



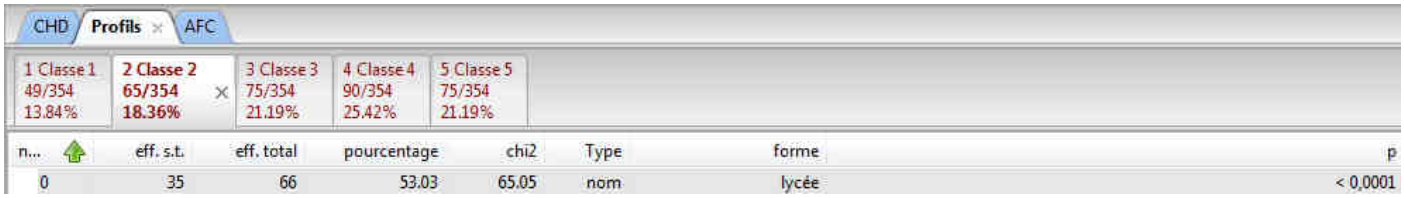
Ce diagramme fournit la liste des formes les plus associées pour chaque classe.

A noter qu'une forme peut se retrouver dans plusieurs classes différentes (voir par exemple 'élève' dans ce cas). Une classe est un regroupement de segments de texte qui contiennent des formes. Le graphique ci-dessus facilite le repérage des formes et leur degré de dépendance aux classes en lisant du haut vers le bas et non l'inverse.

[17] Le deuxième schéma (avec des classes « 0 ») fournit les classes qui ont été abandonnées car ne synthétisant pas assez de données (pour un seuil minimum choisi par défaut cf. p. 18 de ce document.

[18] Il est possible que les classes n'apparaissent pas exactement dans cet ordre selon votre environnement numérique

L'onglet « Profil » est très important pour comprendre le sens des classes :



La classe 2 regroupe 65 segments sur les 354 classés soit (65/354) 18, 36% des segments classés.

Les données sont classées par num (numéros) qui correspondent à un classement décroissant par valeur du Chi².

Il est possible de classer différemment les données en cliquant sur les intitulés de colonnes. Pour revenir au classement initial, cliquer sur la première colonne (num).

L' « effectif s.t. » correspond au nombre de segments comprenant au moins une fois la forme[19]. Par exemple, 35 segments de texte contiennent le mot « lycée » pour la classe 1. L' « effectif total » correspond au nombre total de segments qui contiennent cette forme dans toutes les classes (y compris les classes éliminées lors de la classification). Ces effectifs peuvent être différents de l'effectif de la forme si la forme apparaît plusieurs fois dans un segment. Le pourcentage est calculé par le ratio effectif s.t. sur eff. total ici : 35/66 = 53, 03%

Attention, ceci ne veut pas dire que la forme « lycée » est présente dans 53, 03% des segments du corpus mais que 53,03% des segments contenant la forme « lycée » sont associés à cette classe. En d'autres termes, la majorité des segments contenant la forme « lycée » sont associés à cette classe.

La colonne « **chi²** » donne le résultat du test de dépendance : plus le chi² est élevé, plus l'hypothèse de dépendance entre la forme et la classe est vraisemblable. Selon cette méthode, pour un risque d'erreur de 5% une valeur du chi² théorique de 3, 84 permet de valider la dépendance de deux variables (voir annexe pour le calcul de cette valeur). Dans cette situation, la classe 2 est associée significativement aux mots de numéro 0 (lycée) à 78 (lycéen) qui a un chi² de 4,08. La forme « public », rang 79 avec un chi² de 3,64[20] n'est pas associée significativement à cette classe (Iramuteq fait apparaître NS (Non Significatif) à côté de la p-value pour cette forme). Cette classe est ainsi formée d'environ 80% de formes significativement associées avec un risque d'erreur de 5%.

Enfin, 'p' est la probabilité, le risque que le test de dépendance du Chi² soit faux : plus il est bas, plus la marge d'erreur du test de dépendance est faible. Retenir une valeur de p à 5% est un postulat qui dépend aussi du contexte d'interprétation. Dans une étude sur la perception des effets d'un médicament, le risque toléré ne sera pas le même que pour l'analyse d'un discours par exemple.

En résumé, ce tableau fournit la liste des mots significativement associés à cette classe selon la valeur du chi², de façon plus précise que les dendogrammes. Un chi² supérieur à 10, 827 fournit des résultats très fiables.

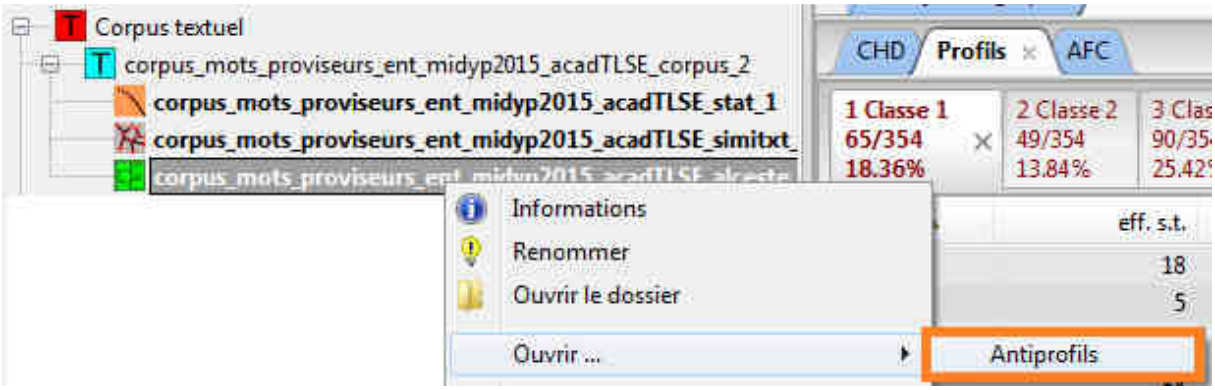
Le tableau de la page suivante est obtenu à partir du fichier « profiles.csv » et permet de résumer les mots ainsi retenus. Il peut servir de base à l'interprétation des données. Mais cette analyse peut aussi être faite directement à partir des données affichées par le logiciel.

[19] Cf. documentation du logiciel p. 24

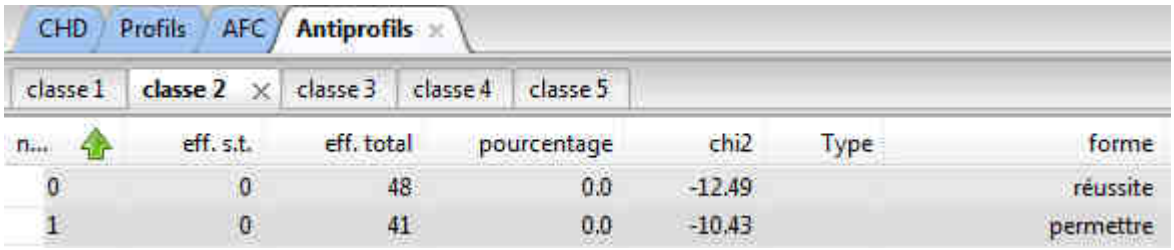
[20] Pour une probabilité d'erreur de 10% (p-value), la valeur du chi² théorique est de 2.706 pour un degré de liberté de 1 (tableau avec une colonne (classe) et une ligne (forme) dans ce cas, voir annexe). Pour un chi² supérieur à **10, 827**, la marge d'erreur est de 0,1%. Le nombre de mots à interpréter par classe est aussi limité par les capacités d'analyse du chercheur… Dans ce dernier cas, le nombre de mots est tout de même de 30 pour la classe 1 avec un chi² supérieur à 10, 827.

	Classe 1			Classe 2			Classe 3			Classe 4			Classe 5		
	Chi²	Forme	P-value	Chi²	Forme	P-value	Chi²	Forme	P-value	Chi²	Forme	P-value	Chi²	Forme	P-value
1	51,21	préparer	8,29E-13	65,05	lycée	7,31E-16	34,47	collège	4,33E-09	61,04	espace	5,58E-15	36,95	valeur	1,21E-09
2	38,71	formation	4,92E-10	36,03	métier	1,95E-09	30,55	année	3,25E-08	43,44	ent	4,37E-11	30,45	confiance	3,43E-08
3	37,99	conduire	7,11E-10	34,98	section	3,33E-09	25,35	difficulté	4,78E-07	42,62	parent	6,64E-11	28,75	respect	8,25E-08
4	37,24	au_delà	1,05E-09	30,64	polyvalent	3,10E-08	24,37	cadre	7,95E-07	42,43	numérique	7,34E-11	26,57	citoyenneté	2,55E-07
5	31,57	supérieur	1,93E-08	26,1	bâtiment	3,24E-07	21,32	humain	3,89E-06	42,43	information	7,34E-11	25,11	éducation	5,42E-07
6	27,16	professionnel	1,88E-07	26,1	durable	3,24E-07	18,87	agréable	1,40E-05	40,29	bienvenue	2,19E-10	22,7	agent	1,89E-06
7	25,27	étude	4,98E-07	22,58	professionnel	2,01E-06	15,15	compte	9,93E-05	30,52	communication	3,30E-08	19,4	école	1,06E-05
8	25,18	post	5,22E-07	22,55	économique	2,05E-06	15,05	arriver	1,05E-04	25,28	souhaiter	4,96E-07	18,06	futur	2,14E-05
9	25,18	initial	5,22E-07	21,75	ouverture	3,11E-06	15,05	accueillant	1,05E-04	25,28	enfant	4,96E-07	17,81	climat	2,44E-05
10	24,71	insertion	6,65E-07	19,5	développement	1,00E-05	15,05	épanouir	1,05E-04	22,63	temps	1,96E-06	17,71	notamment	2,57E-05
11	20,79	technologique	5,12E-06	19,33	entreprise	1,10E-05	15,05	recrutement	1,05E-04	22,62	communauté	1,97E-06	15,05	santé	1,05E-04
12	18,83	orienter	1,43E-05	19,33	domaine	1,10E-05	15,05	internat	1,05E-04	21,97	outil	2,77E-06	15,05	viser	1,05E-04
13	18,83	tournée	1,43E-05	18,98	enseignement	1,32E-05	14,68	situer	1,27E-04	18,96	site	1,33E-05	15,05	servir	1,05E-04
14	18,83	poursuite	1,43E-05	17,99	langue	2,22E-05	14,68	prendre	1,27E-04	17,9	texte	2,32E-05	15,05	condorcet	1,05E-04
15	18,83	positionner	1,43E-05	17,99	label	2,22E-05	14,2	accueil	1,65E-04	17,9	cahier	2,32E-05	15,05	adolescent	1,05E-04
16	18,83	marché	1,43E-05	17,52	général	2,84E-05	12,25	niveau	4,65E-04	16,64	membre	4,53E-05	14,51	jeune	1,39E-04
17	18,61	étudiant	1,60E-05	13,45	particulièrement	2,45E-04	11,89	place	5,65E-04	14,88	visite	1,15E-04	14,12	socle	1,72E-04
18	18,61	paul	1,60E-05	13,45	gallieni	2,45E-04	11,58	accueillir	6,66E-04	14,88	rubrique	1,15E-04	14,12	exigence	1,72E-04
19	18,61	diplôme	1,60E-05	13,45	paix	2,45E-04	11,26	rénover	7,94E-04	14,37	personnel	1,51E-04	14,12	acquérir	1,72E-04
20	18,11	avenir	2,09E-05	13,45	industrie	2,45E-04	11,26	6e	7,94E-04	13,69	disposition	2,16E-04	13,01	ensemble	3,10E-04
21	17,69	réussite	2,60E-05	13,45	françoise	2,45E-04	11,26	quartier	7,94E-04	13,69	note	2,16E-04	12,25	adulte	4,65E-04
22	14,28	former	1,57E-04	13,45	ancrer	2,45E-04	11,26	personnalisés	7,94E-04	13,69	consulter	2,16E-04	11,44	vivre	7,19E-04
23	13,47	technique	2,42E-04	13,42	seconde	2,50E-04	11,26	issu	7,94E-04	11,87	nécessaire	5,71E-04	11,29	devenir	7,81E-04
24	13,47	résultat	2,42E-04	12,85	art	3,37E-04	11,26	3e	7,94E-04	11,87	navigation	5,71E-04	11,26	républicain	7,94E-04
25	12,69	sti2d	3,68E-04	12,85	diversifier	3,37E-04	10,54	élève	1,17E-03	11,87	hésiter	5,71E-04	11,26	prévention	7,94E-04
26	12,69	résolument	3,68E-04	12,85	sport	3,37E-04	10,51	dimension	1,19E-03	10,8	ressource	1,02E-03	11,26	loi	7,94E-04
27	11,81	compétence	5,88E-04	12,85	science	3,37E-04				10,8	différent	1,02E-03	11,26	droit	7,94E-04
28	11,28	filière	7,83E-04	12,85	réaliser	3,37E-04				10,8	améliorer	1,02E-03	11,26	confier	7,94E-04
29	11,23	continuer	8,06E-04	12,66	technologique	3,73E-04				10,78	scolarité	1,02E-03	10,8	poursuivre	1,02E-03
30				12,08	formation	5,09E-04				10,73	travail	1,06E-03	10,8	priorité	1,02E-03

Un autre outil pour comprendre chaque classe est de repérer les formes dites « **anti-profils** » (dans cette situation, formes dissociées de la classe) qui peuvent permettre de vérifier le sens de la classe par son inverse.



Pour la classe 2, les formes anti-profils sont alors :



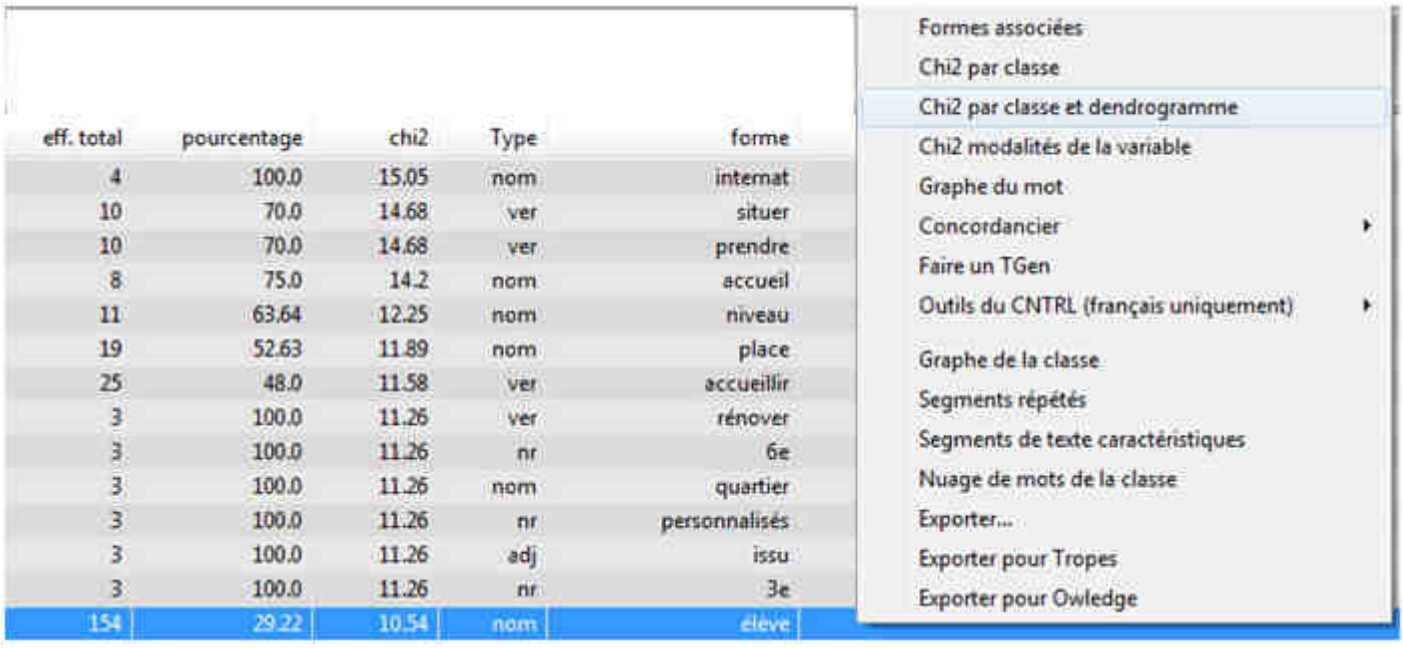
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme
0	0	48	0.0	-12.49		réussite
1	0	41	0.0	-10.43		permettre

Le chi² négatif correspond à la notion d'anti-profil (voir aussi annexe d'explication du chi² 'négatif').

Il existe d'autre façon d'explorer des formes particulières et représenter leur association aux classes.

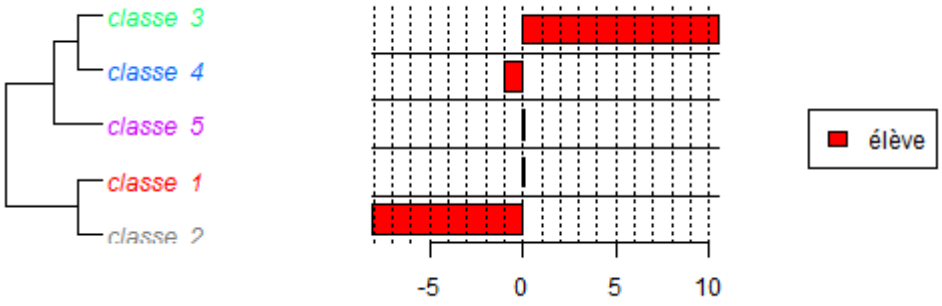
Pour ce corpus, l'analyse de similitude avait montré l'importance du mot « élève » qui n'apparait pourtant pas dans la liste précédente. En effet, ce mot très présent (fréquence maximum) ne semble pas discriminant pour construire une classe. Il est associé à la classe 3 dans cet exemple.

En faisant un clic droit sur cette forme :



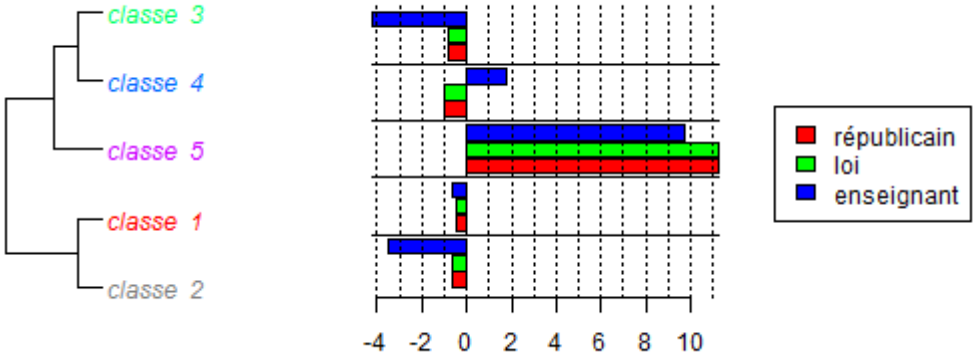
eff. total	pourcentage	chi2	Type	forme
4	100.0	15.05	nom	internat
10	70.0	14.68	ver	situer
10	70.0	14.68	ver	prendre
8	75.0	14.2	nom	accueil
11	63.64	12.25	nom	niveau
19	52.63	11.89	nom	place
25	48.0	11.58	ver	accueillir
3	100.0	11.26	ver	rénover
3	100.0	11.26	nr	6e
3	100.0	11.26	nom	quartier
3	100.0	11.26	nr	personnalisés
3	100.0	11.26	adj	issu
3	100.0	11.26	nr	3e
154	29.22	10.54	nom	élève

On obtient le graphique[21] suivant :



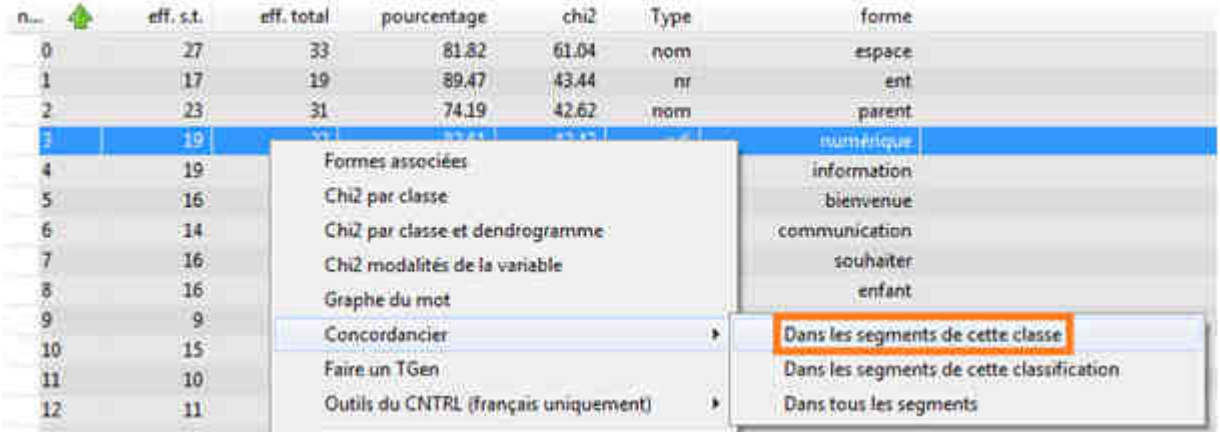
La forme « élève » est associée à la classe 3 et « anti-associée » à la classe 2[22].

Cette même démarche peut être réalisée pour plusieurs mots d'une même classe, exemple : classe 5



Enfin, une dernière méthode (mais pas des moindres) consiste à revenir aux segments de texte.

Par exemple, pour la classe 3 et le mot « numérique », il est possible de retrouver les segments de texte dans lequel il est intégré :



[21] Pour récupérer le graphique, faire une page écran ou utiliser le bouton « Sauver » en bas de la boite de dialogue présentant le graphique.

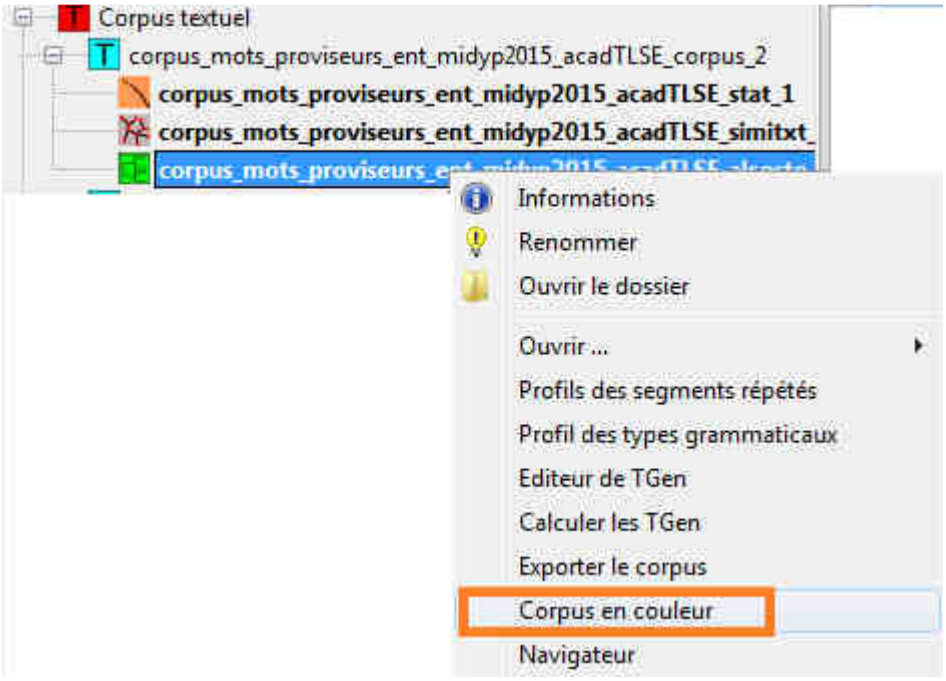
[22] Il est possible que le graphique ait une forme différente cf. remarque p. 19.

On obtient par exemple :

****** *nom_berges *type_lm *dpt_ariege**

nous vous invitons à mieux faire connaissance avec le lycée des métiers du bâtiment bergès grâce à cet environnement **numérique** de travail et en n hésitant pas à nous contacter je vous remercie bonne visite[23]

De la même façon, on peut créer un corpus en couleur qui reprendra l'ensemble du corpus en le colorant selon les classes obtenues par la méthode de Reinert :



Par exemple :

****** *nom_faure *type_lgt *dpt_ariege**

l objectif de ce site est de vous informer sur la vie pédagogique éducative et culturelle de l établissement mais aussi de vous donner l envie d aller à la rencontre de notre communauté scolaire

le lycée gabriel fauré de foix construit à la fin de xix vient d être entièrement rénové par la région c est un beau lycée avec des bâtiments gothiques des jardins une cour d honneur

qui met à la disposition des élèves une diversité d espaces favorables au travail au développement de l autonomie à l enrichissement culturel et au divertissement pour exemples

des box un foyer entre deux jardins 3 salles d étude 3 lieux d exposition l espace deleuze pour des conférences et des projections de films l espace bourdieu réservé aux étudiants post bac

etc le lycée accueille un millier d élèves 190 internes des séries générales et technologiques tertiaires et deux b t s tourisme et gestion

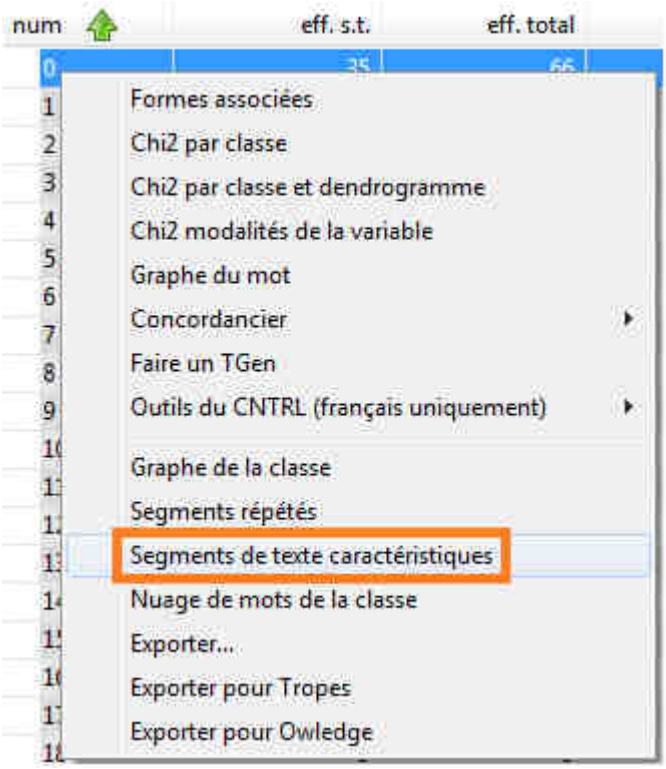
au_delà des résultats le lycée s est donné pour vocation de préparer les élèves à leur accès au post bac ce qui requiert de mobiliser des compétences d adaptabilité de communication et de sérieux dans le travail

pour ce faire dans le cadre de la réforme du lycée une vigilance particulière est accordée aux profils et aux parcours de chacun le lycée s est doté par ailleurs de projets ambitieux dans le domaine du développement durable et de la culture et donne une place effective aux initiatives des lycéens

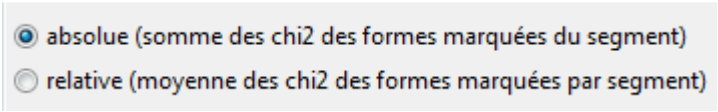
Ce texte comprend 7 segments de texte qui ont été rattachés à 3 classes et un segment non rattaché.

[23] Segment de 40 mots.

Enfin, il est recommandé d'extraire pour chaque classe du corpus quelques segments caractéristiques (clic droit sur une forme de la classe) :



Garder les options[24] par défaut :



Par exemple, pour la classe 2, on obtient :

****** *nom_mathou *type_lp *dpt_hautegaronne**

score : 288.87

le **lycée général** et **technologique** et le **lycée professionnel** ont fusionné au 1er janvier 2015 pour devenir un **lycée polyvalent** le **lycée professionnel** devient une **section** d **enseignement professionnel** rattachée le **lycée polyvalent** conserve son **label lycée** des **métiers** des travaux **publics** et du **bâtiment obtenu** en 2002

L'ensemble de ces méthodes a pour objectif de comprendre le contenu de chaque classe.

[24] La somme des chi² accroit numériquement les différences entre les segments.

ANALYSE FACTORIELLE DE CORRESPONDANCE

L'analyse de correspondance permet de synthétiser l'information et faciliter ainsi son interprétation. Elle est dans ce cas réalisée après une classification de Reinert.

L'algorithme reclasse les formes dans un tableau lexical de ce type[25] :

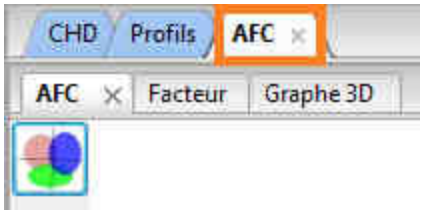
Forme	Classe 1	…	Classe n
Forme 1	1		0
…			
Forme m	0		0

La forme 1 est présente dans un segment de la classe 1.

L'algèbre linéaire permet d'obtenir une représentation en 2 ou 4 dimensions plus facile à analyser au moins pour les deux principales.

Les analyses de correspondances sont ainsi « *largement fondées sur l'algèbre linéaire, produisent des représentations graphiques sur lesquelles les proximités géométriques usuelles entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes* » (Lebart et Salem, 1994 p.80).

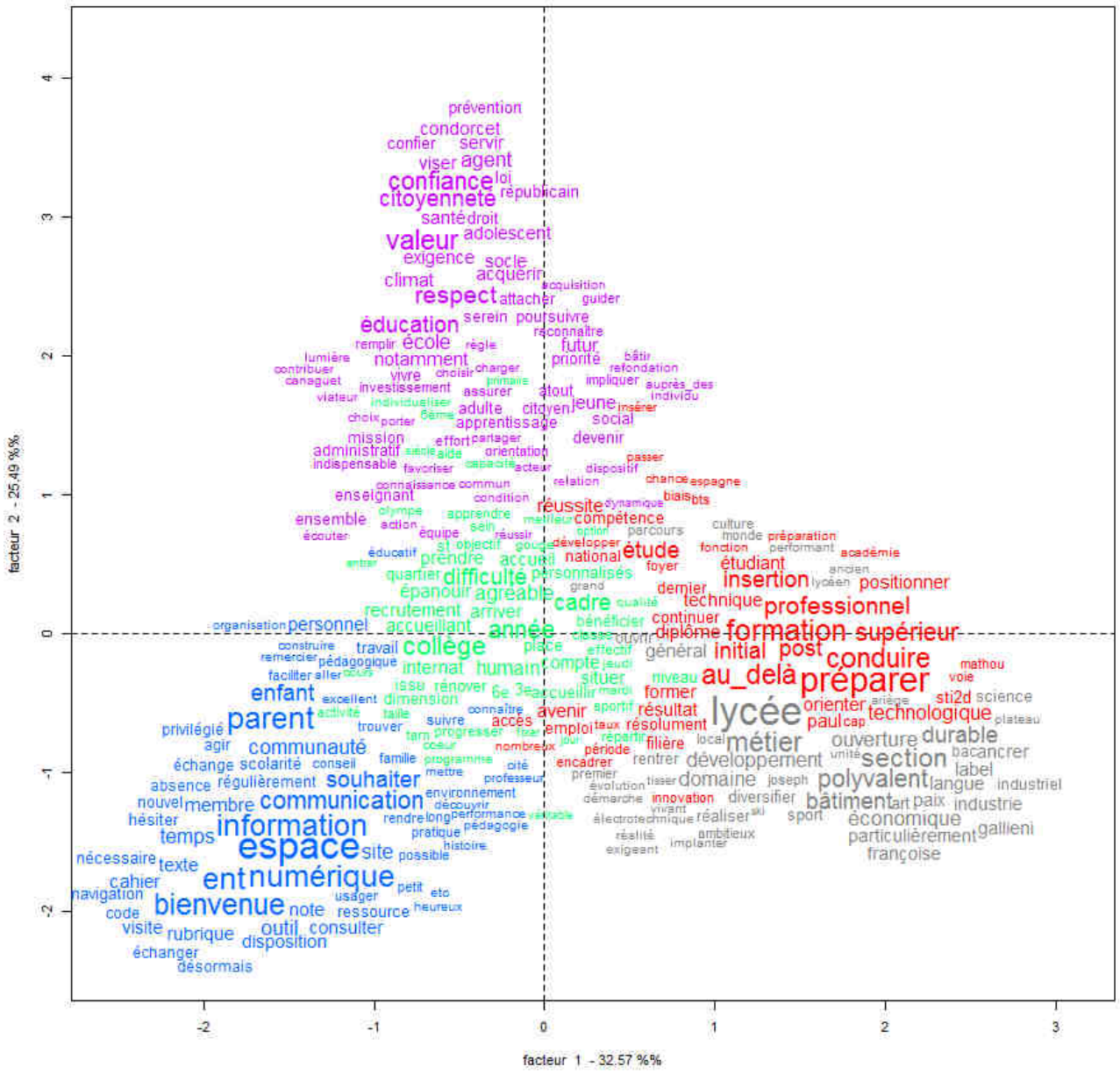
Dans Iramuteq, l'AFC peut être générée[26] à partir de la classification de Reinert à partir de l'onglet AFC :



[25] Dans ce document, l'AFC est créée après la classification de Reinert ce qui explique l'utilisation de classe dans le tableau lexical.

[26] Une autre méthode consiste à analyser les spécificités.

Concrètement, l'AFC va être un graphique avec, par défaut les deux dimensions qui résument le plus de données :



Les formes sont colorées selon les classes de la méthode Reinert. Dans ce cas, les deux premiers facteurs résument 58, 06% de l'information (facteur 1 (32,57%) + facteur 2 (25,49%)).

Des pourcentages plus précis sont fournis dans l'onglet « Facteur » :

AFC | **Facteur** | Graphe 3D

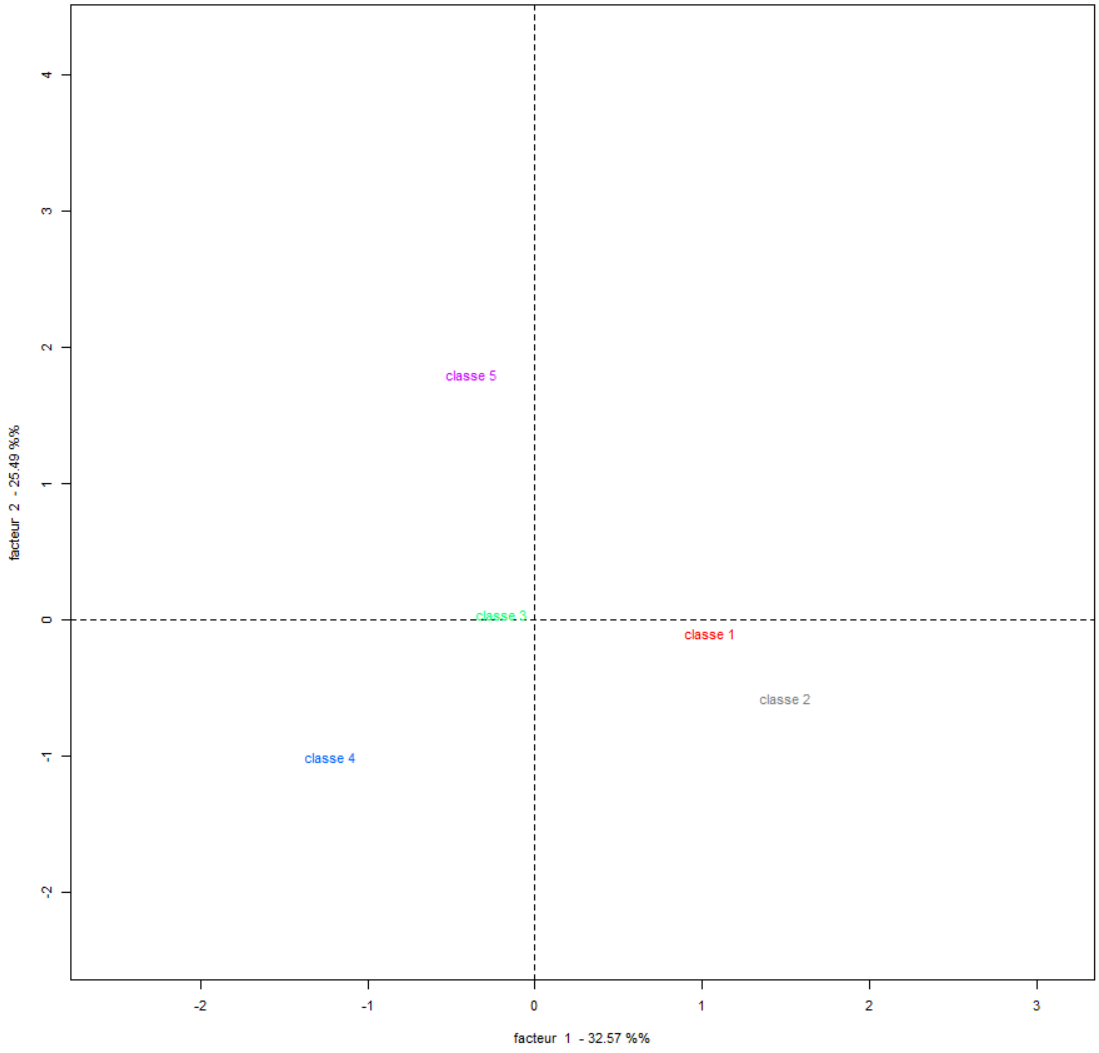
formes	Valeurs propres	Pourcentages	Pourcentage cumules
facteur 1	0.38844	32.56718	32.56718
facteur 2	0.30403	25.48971	58.05689
facteur 3	0.27199	22.80386	80.86075
facteur 4	0.22828	19.13925	100

Un pourcentage inférieur à 20% par exemple indiquerait une AFC peu efficace.

Les couleurs permettent de repérer l'organisation des classes sur ces deux dimensions. Cependant, il peut être possible de simplifier la représentation en indiquant les classes (et non les formes actives) en cliquant sur le bouton suivant :

Le haut de la boite de dialogue suivante doit être configuré de la façon suivante :

Les données représentées sont alors les classes et non plus par défaut, les formes actives. Le graphique de l'AFC est alors :



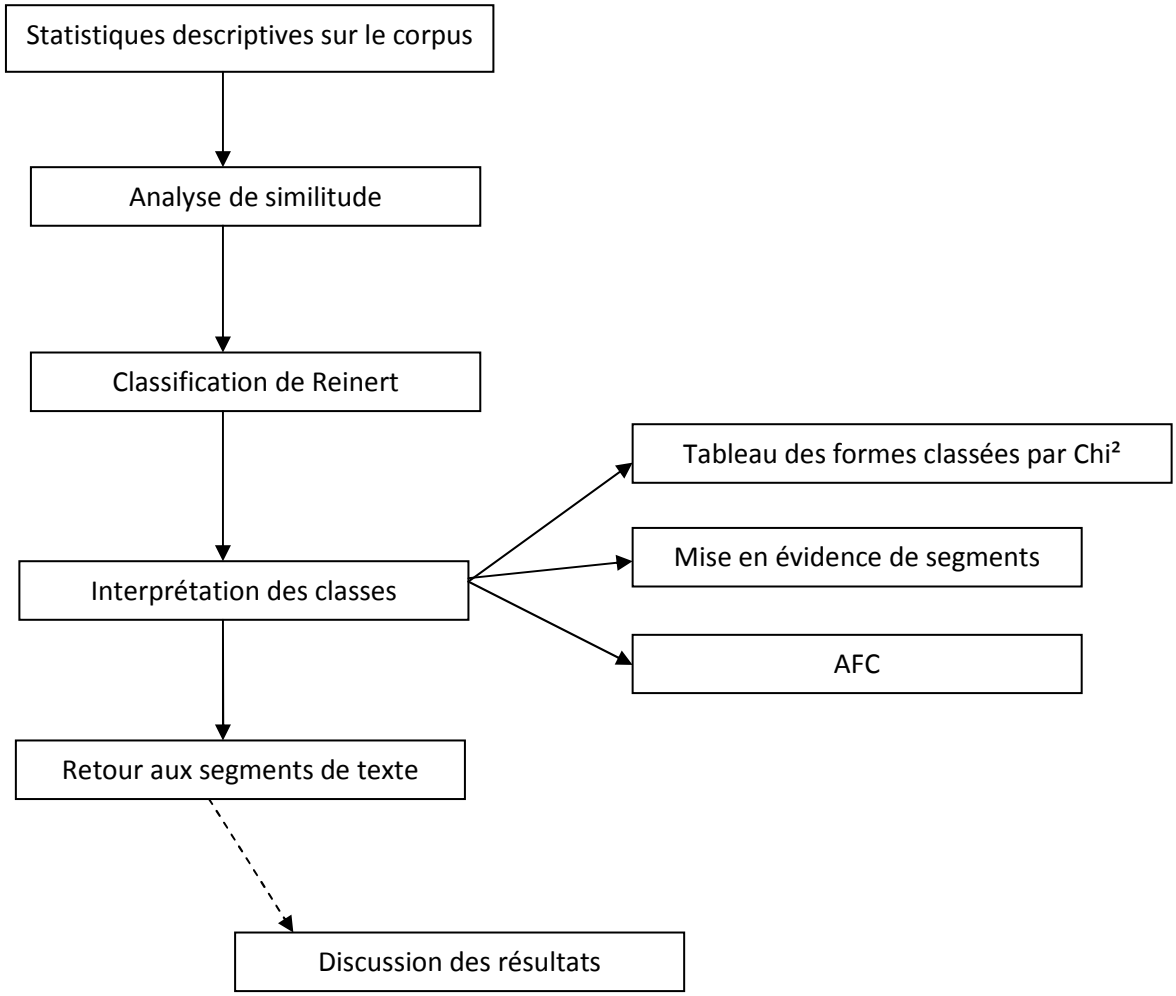
L'interprétation de l'AFC se fait, dans un premier temps, en interprétant les axes par des données opposées. Par exemple, dans ce cas, l'axe vertical oppose les classes 4 et 2. L'axe horizontal oppose les classes 5 et la classe 4. Ces oppositions permettent de nommer les axes et comprendre les résultats de l'AFC. Cette étape est bien-sûr moins systématique et dépend notamment des connaissances et du cadre théorique de l'utilisateur.

Cette interprétation, avec la démarche présentée, se fonde sur la compréhension des classes obtenues par la méthode de Reinert.

Il ne reste alors '*plus*' qu'à discuter les résultats obtenus mais c'est une autre démarche : voir par exemple pour des données *proches* de celles étudiées dans ce document :

http://www.iramuteq.org/Members/dpelissier/analyse-du-discours-d2019etablissements-scolaires-quete-identitaire-et-place-des-ent

Résumé de la démarche présentée



Webographie

Documentation du logiciel :

http://www.iramuteq.org/documentation

Forum du logiciel et mailing list :

https://sourceforge.net/p/iramuteq/discussion/1068065/

Pas à Pas Iramuteq, Elodie Baril et Bénédicte Garnier, Institut National d'Etudes Démographiques Paris (France)

http://www.iramuteq.org/documentation/fichiers/Pas%20a%20Pas%20IRAMUTEQ_0.7alpha2.pdf

Téléchargement du livre Lebart et Salem 1994 :

http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html

Associations spécialisées sur le thème de la lexicométrie :

http://lexicometrica.univ-paris3.fr/

http://www.jadt.org/

Lexicométrie et Alceste :

http://www.ceped.org/IMG/pdf/appliquer_les_methodes_de_la_statistique_textuelle-.pdf

Méthode de classification de Reinert :

https://eudml.org/doc/88079

Le blog d'H. Piment (préparation du corpus notamment) :

http://www.helenepiment.fr/pof/tag/iramuteq/

Bibliographie

BARATS, LEBLANC et FIALA 2013 : Christine Barats, Jean-Mars Leblanc et Pierre Fiala, « *Approches textométriques du web, corpus et outils* », dans Christine Barats (dir.), Manuel d'analyse du web, Armand Colin, Paris p. 100-124

HUSSON, LE et PAGES 2016 : François Husson, Sébastien Lê et Jérôme Pagès, Analyses de données avec R, Presses Universitaires de Rennes.

LEBART et SALEM 1988 : Ludovic Lerbart et André Salem, Analyse statistique des données textuelles, Dunod, 1988.

LEBART et SALEM 1994 : Ludovic Lerbart et André Salem, Statistique textuelle, Dunod, 1994.

MARCHAND 2013 : Pascal Marchand, « Quelques traces chronologiques de l'exploration textométrique », Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique, 120, 38-46, 2013.

RATINAUD 2015 et MARCHAND : Pierre Ratinaud et Pascal Marchand, « Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014) », Mots. Les langages du politique, vol. 108, no. 2, pp. 57–77, 2015.

ANNEXE CALCUL DU CHI² (Laurent Wehrlé)

Chi² pour la forme « lycée »

Cet exemple reprend les données du cas présenté pour la forme « lycée » (classe 2, Chi²=65,05).

Tableau des données observées (voir classe 2 dans Iramuteq) :

Forme	Classe 2	Autres classes	Total
Lycée	35	31	66
Pas lycée	30	258	288
Total	65	289	354

Tableau des données avec hypothèse de dépendance :

Forme	classe 2	Autres classes	Total
Lycée	12,12	53,88	66
Pas lycée	52,88	235,12	288
Total	65	289	354

Calcul de la distance entre données observées et données théoriques :

	classe 2	autres classes
lycée	43,20	9,72
pas lycée	9,90	2,23

Chi²=somme des distances=65,05

La p-value pour un degré de liberté de 1 [(2-1)*(2-1)] est de 7,31E-16 ce qui est une probabilité très faible que la forme 'lycée' ne soit pas associée à la classe 2.

Chi² pour la forme « réussite»

Anti-profil classe 2 (Chi²=-12,49)

Tableau des effectifs observés :

Forme	classe 2	autres classes	total
réussite	0	48	48
pas réussite	65	241	306
total	65	289	354

Tableau des données avec hypothèse de dépendance :

Forme	classe 1	autres classes	Total
réussite	8,81	39,19	48
pas réussite	56,19	249,81	306
Total	65	289	354

Calcul de la distance entre données observées et données théoriques :

	classe 1	autres classes
réussite	8,81	1,98
pas réussite	1,38	0,31

Chi²=somme des distances=12,49

La p-value pour un degré de liberté de 1 [(2-1)*(2-1)] est de 0,000409 ce qui est une probabilité très faible que la forme 'réussite' ne soit pas associée à la classe 2.

En principe, un Chi², qui est le carré d'un écart, est toujours positif. Mais, ici, le Chi² est donné négatif car la dépendance fait apparaître une corrélation négative.