Analyse lexicale appliquée à une question ouverte à l'aide d'IRaMuTeQ

Chloé Six

AGROCAMPUS OUEST

☐ CFR Angers

☑ CFR Rennes

CRÉDOC
CENTRE DE RECHERCHE POUR L'ÉTUDE ET L'OBSERVATION DES CONDITIONS DE VIE

Année universitaire : **2018 – 2019** Spécialité : **Agronome** Spécialisation (et option éventuelle) : **Sciences des données**	**Mémoire de fin d'études** ☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☑ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ d'un autre établissement (étudiant arrivé en M2)

Analyse lexicale appliquée à une question ouverte à l'aide d'IRaMuTeQ

Par : Chloé SIX

Soutenu à Rennes, le 4 septembre 2019

Devant le jury composé de :

Président : **François HUSSON**

Maître de stage : **Gabriel TAVOULARIS**

Enseignant référent : **François HUSSON**

Autres membres du jury : **Sébastien LÊ, enseignant**

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST

REMERCIEMENTS

Je remercie Pascale HEBEL, directrice du pôle « Consommation & Entreprise » du Centre de Recherche pour l'Etude et l'Observation des Conditions de Vie (CRÉDOC) de m'avoir accueillie au sein de son équipe et accompagnée tout au long de ce stage de fin d'études.

Je remercie également Gabriel TAVOULARIS, pour sa tutelle et son encadrement au quotidien ainsi que tous les membres de l'équipe du pôle « Consommation & Entreprise », que j'ai désormais intégré, pour leur constante disponibilité, leur bonne humeur et l'accueil qu'ils m'ont réservé.

Je remercie, de même, François HUSSON, mon tuteur de stage, ainsi que David CAUSEUR, pour leur disponibilité et pour le temps qu'ils m'ont accordé tout au long de ce stage.

Je remercie enfin l'UP mathématiques appliquées pour cette formation et pour chaque instant passé à nous accompagner durant ces trois années.

TABLE DES MATIERES

GLOSSAIRE ET LISTE DES ABREVIATIONS

Corpus	Ensemble de texte
Hapax	Forme citée une unique fois par l'ensemble des individus étudiés
Verbatim	Transcription mot pour mot d'un discours

AFC	Analyse Factorielle des Correspondances
AFDM	Analyse Factorielle de Données Mixtes
ALCESTE	Analyse des Lexèmes Co-occurrents dans les Enoncés Simples d'un Texte
CAH	Classification Ascendante Hiérarchique
CCAF	Comportements et Consommations Alimentaires des Français
CDH	Classification Descendante Hiérarchique
CRÉDOC	Centre de Recherche pour l'Etude et l'Observation des Conditions de Vie
CSP	Catégorie socioprofessionnelle
DGCCRF	Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes
DGE	Direction Générale des Entreprises
IRaMuTeQ	Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires
UC	Unité de Contexte

LISTE DES ILLUSTRATIONS

INTRODUCTION

J'ai effectué mon stage de fin d'études au CRÉDOC (www.credoc.fr), Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie, et plus particulièrement au sein du pôle "Consommation & Entreprise" sous la direction de Pascale HEBEL et sous la tutelle de Gabriel TAVOULARIS.

Le CRÉDOC est un organisme d'étude et de recherche en sciences humaines et sociales au service des acteurs de la vie économique et sociale. Le département « Consommation & Entreprise » s'appuie sur sa connaissance des marchés et son expertise des comportements de consommation pour relier concrètement les stratégies marketing des entreprises aux systèmes d'arbitrages, aux motivations et aux représentations des consommateurs. Le CRÉDOC comprend une trentaine de collaborateurs aux compétences pluridisciplinaires (statisticiens, sociologues, spécialistes du marketing, économistes, linguistes, politistes, ingénieurs agronomes, démographes et nutritionnistes) répartis en 2 pôles d'étude et de recherche, « Consommation & Entreprise » et « Évaluation & Société ».

Le CRÉDOC, association loi du 1er juillet 1901, assure, sous la tutelle du Ministère de l'éducation nationale, de la jeunesse et de la vie associative, de la DGCCRF (Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes) et de la DGE (Direction Générale des Entreprises), des missions d'intérêt général, financées par une subvention de l'État dont le champ et les modalités de prise en charge sont définis par une convention bipartite entre le bénéficiaire et l'organisme financeur. Ces subventions peuvent venir des administrations françaises, des services publics et des organismes paritaires : ministères et assemblées, établissements publics nationaux, agences nationales, grandes caisses de sécurité sociale, collectivités territoriales. Néanmoins, ce financement public ne couvre qu'une part relativement faible des charges. Le CRÉDOC doit assurer l'essentiel de ses recettes par des prestations pour lesquelles il entre en concurrence avec d'autres opérateurs intervenant sur le marché. Il travaille ainsi contractuellement avec des entreprises privées, des banques, des assurances, des fédérations professionnelles, etc.

Depuis sa création en 1953, le CRÉDOC analyse et anticipe le comportement des individus dans leurs multiples dimensions. Celles-ci pouvant être les dimensions « consommateurs », ou « agents de l'entreprise », ou encore « acteurs de la vie sociale ». Il a mis en place depuis 1978 un dispositif permanent d'enquêtes sur les modes de vie, opinions et aspirations des Français et s'est spécialisé dans la construction de systèmes d'information, les enquêtes quantitatives *ad hoc*, les enquêtes qualitatives par entretien ou réunions de groupe et l'analyse lexicale. Le CRÉDOC a été pionnier, dès la fin des années 70, dans l'étude des représentations sociales par le prisme de l'analyse des questions ouvertes et de l'analyse lexicale.

Lors de mon arrivée au CRÉDOC, j'ai eu la responsabilité de réaliser un cahier de recherche, intitulé « La Simplicité Volontaire » sur la base de la question ouverte : « Si je vous dis, « être heureux », à quoi pensez-vous ? ».

Un cahier de recherche bénéficie d'un financement de l'état au titre de la subvention recherche attribuée au CRÉDOC. Il comprend une partie bibliographique sur le thème abordé, une partie contextualisant le sujet, une partie méthodologique, une partie de présentation des résultats, une discussion et enfin une conclusion sur le sujet. Il est réalisé à l'aide de divers corps de métier : sociologues, statisticiens, économistes entres autres. Deux cahiers de recherche ont déjà été réalisés sur cette question : le premier en 1993, dont les données restent introuvables, et le second en 2013.

PROBLEMATIQUE : PRISE EN MAIN DE IRAMUTEQ ET ANALYSE DE L'EVOLUTION DU DISCOURS ENTRE 1993 ET 2019

Comme l'ont montré les trois précédents cahiers de recherche et les travaux du CRÉDOC sur la consommation durable (SIOUNANDAN et al., 2013) (Lahlou et al., 1993) (Sessego and HEBEL, 2018), les comportements sont contradictoires chez les plus diplômés et les plus riches. Ils souhaitent préserver la planète mais ne peuvent pas se passer du confort dans leur logement ou de leur mobilité. Les consommateurs ne peuvent pas toujours avoir le comportement de consommation qu'ils souhaitent à cause de leur mode de vie. En effet, celui-ci a tendance à bousculer leurs habitudes de consommation (Sessego and HEBEL, 2018). En 2013, en fin de crise économique, le CRÉDOC avait mis en avant l'établissement d'une « frugalité choisie », associée à une évolution très importante des représentations du bonheur. Six ans plus tard, dans un contexte social difficile, la mission est d'analyser l'évolution des représentations du bonheur et son association avec des modes de consommations frugaux. Elle sera aussi de confronter les représentations mentales du bonheur aux nouvelles pratiques de consommation tout en déterminant les facteurs sociologiques les plus explicatifs. Tout ceci se fera à l'aide des questions de l'enquête « Tendances de consommation » du CRÉDOC. Cette enquête est effectuée depuis 30 ans et porte sur les arbitrages de consommation, les critères d'achat, les nouveaux modes de consommation et les choix environnementaux. Elle est réalisée auprès d'un échantillon représentatif des adultes de 18 ans et plus de la population française.

L'objectif de mon stage est donc d'étudier l'évolution des représentations du bonheur et son association aux tendances de consommations de 1993 à 2019. Je ne présenterai que la partie qui concerne les représentations du bonheur, le reste de l'analyse fera l'objet d'une publication par le CRÉDOC d'ici la fin de l'année 2019 sur son site internet. Ne pouvant avoir accès aux données de 1993, je me contenterai seulement des résultats publiés (SIOUNANDAN et al., 2013) (Lahlou et al., 1993). L'enquête « Tendances de consommation » débutant en juin, j'ai pu effectuer toutes les étapes de celle-ci : création/rédaction du questionnaire, gestion du terrain d'enquête, calcul des quotas, nettoyage de la base (arrivée mi-juillet) sous SAS et son analyse sous IRaMuTeQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires, iramuteq.org) et R.

Les analyses lexicales étaient, à l'époque (1993 et 2013), faites sur le logiciel ALCESTE (Analyse des Lexèmes Co-occurrents dans les Enoncés Simples d'un Texte). Ce logiciel étant payant, il a été recodé par Pierre Ratinaud et Pascal Marchand et mis à disposition gratuitement sous le nom d'IRaMuTeQ. Le CRÉDOC a donc décidé d'abandonner ALCESTE pour utiliser IRaMuTeQ. Il m'a été assigné comme mission d'appréhender ce nouveau logiciel, de comprendre les analyses qu'il propose et les paramètres associés à ces analyses afin de transmettre ma compréhension à l'équipe. J'ai donc comparé les deux logiciels sur les données de 2013, et ai cherché à comprendre les méthodes utilisées (AFC, chi2, classification descendante hiérarchique selon la méthode de Reinert) ainsi que les paramètres d'IRaMuTeQ.

Dans ce rapport, je me poserai la question de savoir quelles analyses offre le logiciel IRaMuTeQ. Je me demanderai aussi comment l'on peut décrire et comprendre la méthode de Max Reinert (1983) – appelée classification descendante hiérarchique. Enfin, je me focaliserai sur comment évaluer l'évolution du discours entre 1993 et 2019, et quelle est-elle ?

Après avoir passé en revue les particularités et les objectifs de l'analyse lexicale, j'aborderai la description des différentes analyses proposées par IRaMuTeQ avec une étude approfondie de la méthode de Reinert (1983) – la classification descendante hiérarchique – ainsi que ses paramètres d'entrée. Enfin, je traiterai le sujet du cahier de recherche en observant comment comparer l'évolution des discours et son vocabulaire entre 2013 et 2019.

I. L'ANALYSE LEXICALE

L'analyse lexicale est une façon d'analyser le langage. Appliquée au verbatim des consommateurs, elle permet d'obtenir les représentations que se font les individus d'un objet à partir des mots associés (LAHLOU, 1993). L'analyse de données textuelles, inspirée par la linguistique structurelle et l'analyse de discours, est autant qualitative que quantitative. Elle cherche à qualifier les éléments des textes à l'aide de catégories et à les quantifier en analysant leur répartition statistique. Cette approche s'inspire principalement des travaux de Jean-Paul Benzécri (1960) (Beaudouin, 2016).

A) OBJECTIF DE L'ANALYSE LEXICALE

L'analyse lexicale, selon Saadi LAHLOU, est « une approche des sciences humaines qui envisage les textes comme des données organisées. C'est l'art d'extraire et de synthétiser les concepts abordés dans un corpus » (LAHLOU, 1993). Elle permet d'accéder aux représentations mentales d'un objet ou selon Max Reinert (1993) à des « mondes lexicaux » par une analyse approfondie du discours spontané des enquêtés. Cela se fait en décrivant les individus par les concepts qu'ils ont cités et non par le sens de leurs réponses. Elle met à jour la variabilité des discours par la description typologique du corpus à l'aide des liens entre les individus et la mise en évidence des proximités sémantiques (LAHLOU, 1992).

Le principe est de structurer les données textuelles en un tableau « individus x variables » sur lequel seront appliquées des techniques d'analyse de données. Ces méthodes multivariées sont exploratoires, c'est-à-dire que les données sont observées sans *à priori*. L'analyse factorielle des correspondances (AFC), mise au point à partir des années 1960 par Jean-Paul Benzécri, est la méthode de prédilection en analyse des données textuelles (BENZECRI, 1973). Cette méthode permet de rendre compte graphiquement de la proximité entre des individus qui ont le même profil de vocabulaire d'une part et entre des mots utilisés par les mêmes individus d'autre part. C'est sur cela que le CRÉDOC a formalisé et développé, à la fin des années 1980, des méthodes qui tiennent compte à la fois d'une approche statistique du lexique et d'une approche psychosociologique. L'unité statistique étudiée n'est plus le mot mais la réponse entière de l'individu à une question ouverte. En rapprochant les énoncés qui utilisent le même lexique, les « mondes lexicaux » qui constituent les représentations sociales sont mis en lumière au moyen d'une classification descendante hiérarchique (Reinert, 1993).

B) LA FORMULATION DE LA QUESTION OUVERTE

Il est important de noter la différence entre une question fermée et une question ouverte. Une question fermée est par définition une question dont la réponse est à donner parmi un choix de réponses limité. Le choix peut être unique comme multiple. Une question ouverte n'a pas de réponse préétablie, la réponse est libre, c'est au panéliste de trouver lui-même une réponse (Lahlou, 1993). La question ouverte permet d'obtenir des avis ou plus précisément des représentations mentales d'objet. Pour cela, il est nécessaire de poser la question de manière particulière afin de récolter l'information souhaitée (LAHLOU, 1992). Il est donc important de poser une question claire au panéliste, qui sera interprétée de la même façon par l'ensemble de la population interrogée. Il est aussi essentiel que la question ouverte soit la première posée pour éviter qu'une des questions précédentes ne biaise ou n'influence la réponse de la personne interrogée (LAHLOU, 1992).

Pour chaque individu, la représentation qu'il donne de l'objet est subjective, il est difficile de la ramener à des critères objectifs. Il faudrait rassembler plusieurs disciplines pour comprendre les différents aspects de la représentation sociale (Figure 1) (LAHLOU, 1992).

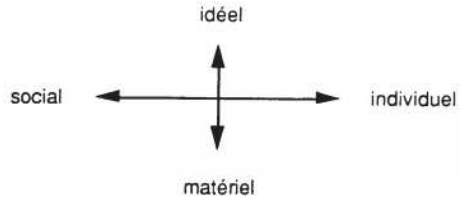

Figure 1 : La représentation sociale vue comme un carrefour (LAHLOU, 1992)

Saadi Lahlou suppose que les représentations fonctionnent suivant le modèle « *si... alors* » (LAHLOU, 1992). L'idée est de donner à l'individu le « *si* » pour avoir son « *alors* ». Il pense qu'après le « *si* », le « *alors* » vient naturellement, par simple évocation. Il est aussi capital, dans le cadre des études du CRÉDOC d'impliquer la personne interrogée dans la question, afin d'avoir son opinion et non une vérité générale ou un stéréotype, d'où le pronom « vous » dans la question « Si je *vous* dis ». On récupère alors soigneusement les « évocations obtenues chez le sujet » (LAHLOU, 1992). Saadi Lahlou ajoute que le choix du libellé de la question est délicat car « choisir, c'est renoncer à certains aspects qui auraient pu être apportés par une autre formulation » (LAHLOU, 1992).

Il est difficile de savoir si les réponses données par le sujet sont personnelles ou générales. Pour l'individu, il n'existe pas de différence entre représentation individuelle et représentation sociale, il la fait fonctionner au même titre que les « *si ... alors* » qui proviennent de son expérience personnelle individuelle (LAHLOU, 1992). Il est également délicat de percevoir ce qui ressort réellement des représentations du sujet une fois qu'il les évoque de manière écrite ou orale. C'est-à-dire que celui-ci peut, intérieurement, avoir une image ou un ressenti très précis mais qui, décrit par ses mots, peut perdre en clarté ou en signification.

Deux nouveaux problèmes apparaissent donc : celui de la relation entre représentation sociale et représentation mentale, et celui du rapport entre représentation mentale et le langage. C'est grâce à l'analyse que l'on va reconstruire les représentations mentales qui correspondent à la représentation sociale.

Pour résumer, la philosophie générale de l'analyse (Yvon, 1990) (LION, LAHLOU, 1991) consiste à :

- Obtenir une image de l'objet dont on cherche à connaître la représentation sociale en recueillant un corpus de phrases « à propos de cet objet » ;
- Reconstruire à partir de ce corpus un espace d'association dans lequel figurent les mots ou les racines utilisés.

Sur le plan technique, l'opération se segmente en 4 phases (Figure 2) :

- Recueil et construction du corpus ;
- Traitement du corpus (SAS, R, IRaMuTeQ) ;
- Analyse des résultats ;
- Interprétation.

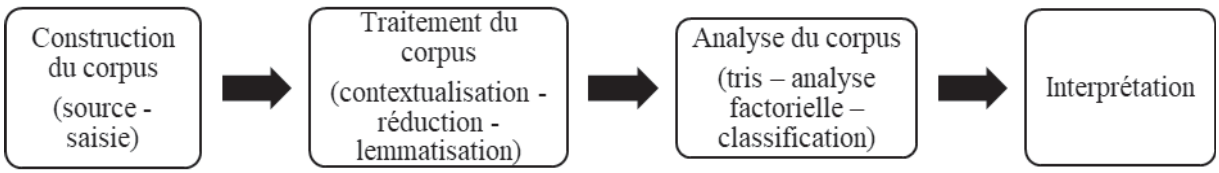

Figure 2 : Les quatre moments de l'analyse (LION, LAHLOU, 1991)

Dans ce rapport, le but de la question ouverte est d'accéder à la représentation mentale qu'ont les Français de la notion « être heureux » et de trouver les différentes dimensions de sa représentation sociale. Pour cela, le CRÉDOC utilise la méthode développée par Reinert (1983), recodée aujourd'hui dans un nouveau logiciel libre: IRaMuTeQ.

II. IRAMUTEQ ET SES METHODES STATISTIQUES

IRaMuTeQ permet de faire des analyses sur des corpus de texte. Une première analyse propose des statistiques simples sur les corpus textuels : effectif des formes actives et supplémentaires et liste des hapax. L'analyse suivante s'appelle « Spécificité et AFC ». Cela permet, à l'aide de l'indice du chi2 ou de la loi hypergéométrique, de déterminer les lemmes significativement plus cités par l'une des modalités d'une variable sociodémographique comparativement aux autres (tris croisés). On peut aussi appliquer la méthode de Reinert afin de faire une typologie de discours. On peut également effectuer une analyse des similitudes ou encore créer un nuage de mots qui corrèle la taille des mots avec leurs effectifs relatifs.

Pour comprendre les analyses effectuées dans le logiciel IRaMuTeQ et en particulier, la classification descendante hiérarchique ainsi que les paramètres que l'on a en entrée de celle-ci (Loubère, Ratinaud, 2014) et l'importance de la lemmatisation, j'ai décidé de recoder la méthode de Reinert avec R grâce à ma compréhension des articles trouvés dans la littérature.

A) LA CLASSIFICATION DESCENDANTE HIERARCHIQUE – CDH

1) La méthode Reinert (1983, 1986 et 1990)

La classification descendante hiérarchique (CDH) est une procédure croisant plusieurs techniques. Elle est principalement élaborée en relation avec une pratique d'analyse de données en psychosociologie, avec en particulier l'analyse des réponses libres à des questions ouvertes. La nature des données textuelles d'entrées peut néanmoins être très diverse, il peut ainsi s'agir d'entretiens semi-directifs, de corpus, de résumés d'articles, de corpus de discours politiques, ou encore de romans. Selon Max Reinert (1983, 1987, 1990) : « La procédure descendante de la classification s'explique par le souhait de posséder des classes de cardinaux assez élevés, bien différenciées les unes des autres afin de les décrire correctement ». Le jeu de données initial est un tableau à double entrée avec en lignes tous les individus, et en colonnes toutes les formes (mots) citées. Au croisement « individus x formes » se trouve un indicateur d'absence-présence (soit 0 ou 1) indiquant si la forme a été citée par l'individu ou non. Le nombre total de « présence » est relativement faible ; on appelle cela des tableaux hypercreux.

Suivant la longueur d'une réponse le texte peut être divisé en unités de contexte (UC) (qui devient « individu ») en faisant l'hypothèse que ces UC renvoient à des représentations sous-jacentes que l'on peut expliciter. Chaque UC est de petite taille (entre 1 et 10 lignes). Dans les enquêtes du CRÉDOC, les réponses aux questions ouvertes ne sont pas assez longues pour les diviser (il n'est demandé que 5 mots). De plus, il est préférable qu'un individu ne se trouve que dans une unique classe de la typologie de discours, il ne faut donc pas que sa réponse soit coupée en morceaux, sous peine de voir apparaître ces morceaux dans différentes classes.

Le format d'entrée des données dans IRaMuTeQ est assez particulier et demande un prétraitement (Tableau 1). Il y a d'une part le « *texte* » qui constitue la partie à analyser, pouvant être les réponses à la question ouverte et d'autre part une partie dite « *hors corpus* », c'est-à-dire qu'elle ne fait pas partie de la réponse mais elle peut la caractériser lors de l'analyse. Cette partie peut contenir des caractéristiques qualitatives comme des variables sociodémographiques comme par exemple : le sexe, l'âge, la CSP (catégorie socioprofessionnelle), etc.

Tableau 1 : Illustration par un exemple du format d'entrée des données dans IRaMuTeQ

Variables « *hors corpus* »	**** *ID_1 *SEXE_Homme *AGE_18_24_ans
Texte	Santé, Amour, Famille, Travail

Les deux logiciels reconnaissent les formes dites « *mots-outils* » : articles, prépositions, conjonctions, pronoms, auxiliaires qui ne sont pas compris dans l'analyse principale et qui sont considérés comme des variables supplémentaires. L'utilisateur du logiciel est néanmoins libre de choisir s'il veut les analyser ou non. Il faut garder le plus grand ensemble possible de mots en effectuant tout de même certaines modifications sur les formes brutes pour harmoniser le vocabulaire : supprimer les formes du pluriel, les marques de conjugaison et certains suffixes pour pouvoir conserver un contexte plus large. Par exemple « chantons, chanter, chantonner » donnent « chanter ». On appelle cela la lemmatisation. Cela permet d'avoir le plus de liaisons statistiques possibles impliquées dans les cooccurrences de formes. Il y a certes une perte d'information, mais cette étape est nécessaire si l'on veut que l'information conservée permette une analyse intéressante du tableau croisé hypercreux. Plus les formes sont regroupées et plus elles ont de poids dans l'analyse par leur grand nombre.

La méthode de Reinert (1983) consiste à décrire les lois de distribution du vocabulaire dans les textes. En plus de cela, il s'agit d'étudier les types de représentations au travers de ces lois de distribution. En analysant les ressemblances et dissemblances des vocabulaires, on peut observer des variations entre les formes de relations aux mondes lexicaux (Reinert, 2008). Le but de l'algorithme est de réordonner le tableau en deux sous-ensembles, en retirant des lemmes non caractéristiques (Figure 3).

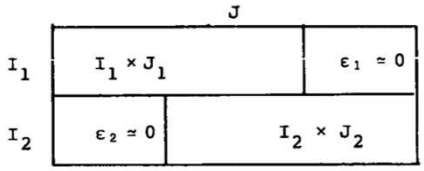

Figure 3 : Tableau souhaité après la première itération de l'algorithme (Reinert, 1983)

Seules les formes analysables (verbes, noms, adverbes et adjectifs) sont utilisées pour obtenir la typologie tandis que les formes illustratives servent uniquement à la description des classes. Les variables « *hors corpus* » sont considérées comme des formes illustratives, elles permettent aussi de caractériser les classes de discours. On aboutit donc à la séparation du jeu de données principal en deux classes chevauchantes (Figure 3). La séparation en deux sous-ensembles se fait en maximisant le moment d'ordre 2 ou chi2 (le moment d'ordre deux est le chi2/N du tableau condensé, N étant le nombre total de mots cités par tous les individus (Reinert, 1983). Une fois la partition des individus obtenue, les lemmes ou formes associés aux deux sous-ensembles (en colonne) sont gardés en fonction d'une valeur liée au chi2 ; au-dessus de cette valeur, le lemme appartient au sous-ensemble le citant le plus, sinon chacun des sous-ensembles le conserve (d'où le nom de classes chevauchantes). La classe ayant la plus grande hétérogénéité (1- C) (Figure 4) est divisée à nouveau (Ratinaud, 2018).

$$c = \chi^2 \times \frac{N}{Nbre\ de\ lignes \times Nbre\ de\ colonnes}$$

Figure 4 : Calcul de l'homogénéité, N étant le nombre d'occurrences totales (Ratinaud, 2018)

La classe ayant à la fois le plus d'individus et le plus de lemmes restants est celle qui est soumise à nouveau à l'algorithme. On a donc construit un ensemble de sous-tableaux emboités, indicé par le chi2 de ces derniers. Le processus est itératif et cesse lorsque l'utilisateur possède le nombre de classe souhaité.

En résumé, on recherche la partition en deux classes maximisant le moment d'ordre deux, puis on recherche les classes chevauchantes des variables associées à cette partition. Même si l'algorithme élaboré ne permet pas d'affirmer que la partition obtenue maximise au mieux le moment d'ordre deux, il l'approche le plus possible.

Dans le détail, comment fonctionne l'algorithme ?

1) Recherche du premier axe factoriel.
    a. Faire une AFC sur le jeu de données initial.
    b. Récupérer les coordonnées ordonnées des individus sur le premier axe factoriel de l'AFC. Cet axe est celui ayant une valeur propre la plus haute donc la plus grande variabilité.
2) Recherche de l'hyperplan orthogonal en découpant le premier axe factoriel en deux en cherchant à maximiser le moment d'ordre deux ou chi2 (Figure 5) des deux sous nuages.

$$Chi^2 = \sum_{i,j} \frac{(Eff.théo._{(i,j)} - Eff.\ obs._{(i,j)})^2}{Eff.théo._{(i,j)}}$$

Figure 5 : Calcul du chi2 (adscience.fr)

3) Amélioration de la partition optimale obtenue : pour chaque individu on observe si sa permutation augmente le moment d'ordre deux ou non. Ainsi, on améliore petit à petit la partition initiale en trouvant un maximum local du moment d'ordre deux.
4) Utilisation du chi2 à un degré de liberté pour comparer, pour chaque lemme sa distribution dans les deux classes. Pour rendre la valeur indépendante du nombre de fois kj que ce lemme a été cité dans l'ensemble des réponses, on crée un nouvel indicateur $C = \sqrt{\frac{\chi^2}{\chi^2+kj}}$.
5) Calcul de l'homogénéité (Figure 4) de chaque sous-tableau et nouvelle soumission à l'algorithme du sous-tableau ayant la valeur d'homogénéité la plus faible.

Plusieurs paramètres existent : CTEST est le seuil du coefficient C pour lequel le recouvrement des classes que l'on trouve est optimal à 0,3 et TJS qui est la fréquence pour laquelle les lemmes ayant été cités moins de fois que celle-ci sont supprimés, la classification est optimale pour TJS valant 3 (Reinert, 1983).

2) Illustration de la méthode de Reinert

Pour que tout ceci soit plus concret et plus facile à comprendre, prenons un exemple simple (Tableau 2).

Tableau 2: Matrice croisant les individus I en lignes et les formes J en colonnes (Ratinaud, 2018)

	J1	J2	J3	J4	J5
I1	0	1	0	0	1
I2	1	1	0	0	0
I3	0	0	1	1	1
I4	0	1	1	1	0
I5	1	1	1	0	1

Ce tableau de données va donc subir une AFC (Figure 6a). Chaque individu est projeté sur l'axe 1 (Tableau 3). On obtient une liste avec les individus triés par leur coordonnée sur le premier axe factoriel de l'AFC.

Tableau 3 : Coordonnées des individus sur le premier axe factoriel de l'AFC

I2	I5	I1	I4	I3
-1,1666153	-0,3411973	-0,2329329	0,5516837	0,8362781

Il faut donc ensuite trouver la partition qui maximisera le chi2. Pour cela, on coupe le premier facteur entre I2 et I5, on calcule alors le chi2 entre la classe 1 contenant I2 et l'autre classe contenant I5, I1, I3, I4. On retient cette valeur de chi2. On sépare ensuite de manière à avoir I2 et I5 dans la même classe et dans l'autre I1, I3, I4. On calcule la valeur du chi2 entre ces deux classes. Ainsi de suite jusqu'à laisser I3 seul de son côté et comparer la classe I3 avec la classe I2, I5, I1, I4. On a donc les valeurs du chi2 pour toutes les partitions, on prend le maximum de cette liste et on retrouve quelle partition donnait cette valeur. Dans notre exemple, la coupure maximisant le chi2 correspond à celle qui sépare les individus se situant entre I1 et I4 (Figure 6b). La partition optimale est donc, pour la première classe I2, I5, I1 et pour l'autre classe I3 et I4.

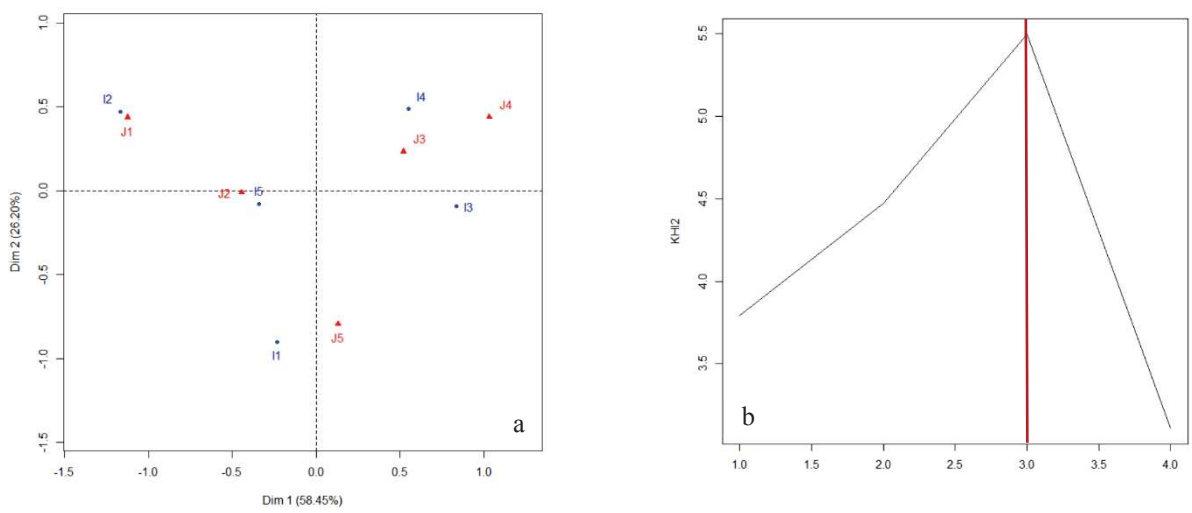

Figure 6 : AFC du tableau 1 (a) et l'évolution du chi2 en fonction de la partition faite (b)

Pour illustrer le calcul du chi2, prenons la séparation optimale (Tableau 4). Nous récupérons pour chaque classe de la partition, le nombre de fois que le lemme j {1,2,3,4,5} a été cité par l'ensemble de la classe. On calcule dans le même temps le nombre de fois théorique que cette classe aurait dû citer chacun des lemmes en faisant :

$$\frac{Marge\ lignes \times Marge\ colonnes}{Somme\ totale\ des\ mots\ cités}$$

Tableau 4 : Tableau de la partition optimale avec les valeurs observées et théoriques

Observé	J1	J2	J3	J4	J5	Marge lignes
C1 {I3, I4}	0	1	2	2	1	6
C2 {I1, I2, I5}	2	3	1	0	2	8
Marge colonnes	2	4	3	2	3	**14**

Théorique	J1	J2	J3	J4	J5
C1 {I3, I4}	(2*6) / **14**= 0,857	(4*6) / **14** = 1,714	1,286	0,857	1,286
C2 {I2, I5, I1}	(2*8) / **14** = 1,143	2,286	1,714	1,143	1,714

La valeur du chi2 vaudra :

$$\frac{(0 - 0{,}857)^2}{0{,}857} + \frac{(1 - 1{,}714)^2}{1{,}714} + \cdots + \frac{(2 - 1{,}717)^2}{1{,}714} = 5{,}4931$$

Ensuite, afin de savoir quel lemme appartient à quelle classe, nous calculons le chi2 pour un lemme précis entre les deux classes et ce, pour chaque lemme.

Pour rendre la valeur indépendante du nombre total d'occurrences du lemme on applique au chi2 :

$$C = \sqrt{\frac{\chi^2}{\chi^2 + marge\ colonnes}}$$

Si cette valeur vaut plus de 0,3 (CTEST), alors le lemme est associé à la classe qui le cite le plus de fois, sinon, il est associé aux deux classes (Tableau 5).

Tableau 5 : Les différentes étapes pour associer les lemmes aux classes

	J1	J2	J3	J4	J5	Marge lignes
Observé	0	1	2	2	1	6
	2	3	1	0	2	8
Marge colonnes	2	4	3	2	3	**14**
Théorique	0,857	1,714	1,286	0,857	1,286	
	1,143	2,286	1,714	1,143	1,714	
Chi2 par lemme	1,500	0,521	0,694	2,667	0,111	
C	**0,655**	**0,339**	**0,434**	**0,756**	0,189	
A quelle classe le lemme appartient-il ?	**2**	**2**	**1**	**1**	**1 et 2**	

La classe la plus hétérogène selon Ratinaud est divisée à nouveau par le même procédé.

B) IMPORTANCE DE LA LEMMATISATION

La lemmatisation est un traitement lexical apporté à un texte pour son analyse. Il s'agit de faire correspondre aux occurrences des lemmes leur forme enregistrée dans le dictionnaire.

1) Les choix arbitraires de modification du dictionnaire de lemmatisation

Après avoir préparé le fichier texte des réponses, il faut l'importer et décider de le lemmatiser avec un dictionnaire classique ou un dictionnaire modifié préétabli auparavant. Le dictionnaire peut être modifié suivant des décisions arbitraires que l'on considère nécessaire. Par exemple « été » sera par défaut lemmatisé en l'auxiliaire « être » alors qu'il peut signifier la saison de l'été.

Pour cela, il faut remplacer dans les données de départ le « été » par « saison_été » lorsqu'il s'agit de la saison de l'« été » et ajouter dans le dictionnaire *« saison_été été nom »* pour faire correspondre le mot « saison_été » à la lemmatisation été se trouvant être un nom. La lemmatisation est très importante pour l'analyse.

De plus, chaque lemme garde son rôle grammatical, à l'inverse d'ALCESTE qui lui, pouvait regrouper deux termes sous la même racine lexicale même s'ils n'appartenaient pas à la même classe grammaticale. Par exemple, profession et professionnellement sont recodés sous le lemme « profession+ » par ALCESTE mais sont deux lemmes différents sous IRaMuTeQ, ce qui, dans le cas des réponses à la question ouverte de l'enquête « Tendances de consommation », est problématique, car l'idée ou le concept reste le même. Que l'on réponde « avoir une profession épanouissante » ou « être épanoui professionnellement », l'idée est de saisir le concept d'« un épanouissement professionnel ». Il faut donc manuellement modifier le dictionnaire afin de pallier cette distinction de classe grammaticale que fait IRaMuTeQ.

2) L'influence de l'AFC

La première étape de la méthode de Reinert est une AFC (1983). L'AFC utilise la métrique du chi2. Or celle-ci introduit les inverses des fréquences marginales des modalités (colonnes) comme pondération des écarts entre éléments de deux profils relatifs (individus), et réciproquement. Elle attribue donc plus de poids aux écarts correspondants à des modalités de faible effectif (rares) sur les colonnes.

Cela signifie que si « enfants » et « enfant » sont cités réciproquement 7 et 3 fois, et que tous les autres mots sont cités 10 fois, « enfant » apparaitra comme une modalité rare et modifiera le nuage de l'AFC. Comme « enfants » et « enfant » correspondent à un même concept relatif au bonheur, on souhaite une unique lemmatisation sous le lemme « enfant » qui permettra à l'AFC de n'avoir aucune modalité rare. On peut voir dans l'exemple ci-dessous (Figure 7), tiré des données de 2019, que si le corpus n'est pas, ou est mal lemmatisé, alors certains mots ayant le même sens sont significatifs pour deux classes, « amour » et « amoureux » ainsi que « libre » et « liberté », ce qui ne représente pas des classes optimales pour nos analyses. En première partie, nous parlions de représentations mentales et de concepts. Dans ce cas précis, l'utilisation du vocabulaire est différente ; il peut varier selon de multiples facteurs (éducation, milieu social, âge, sexe, etc.) mais le concept sous-jacent et donc la représentation mentale du bonheur reste semblable si l'on parle d'être « libre » ou de « liberté ».

Il est important de lemmatiser les mots parlant du même concept afin de le dissocier des autres concepts existants et d'établir les différentes représentations du bonheur.

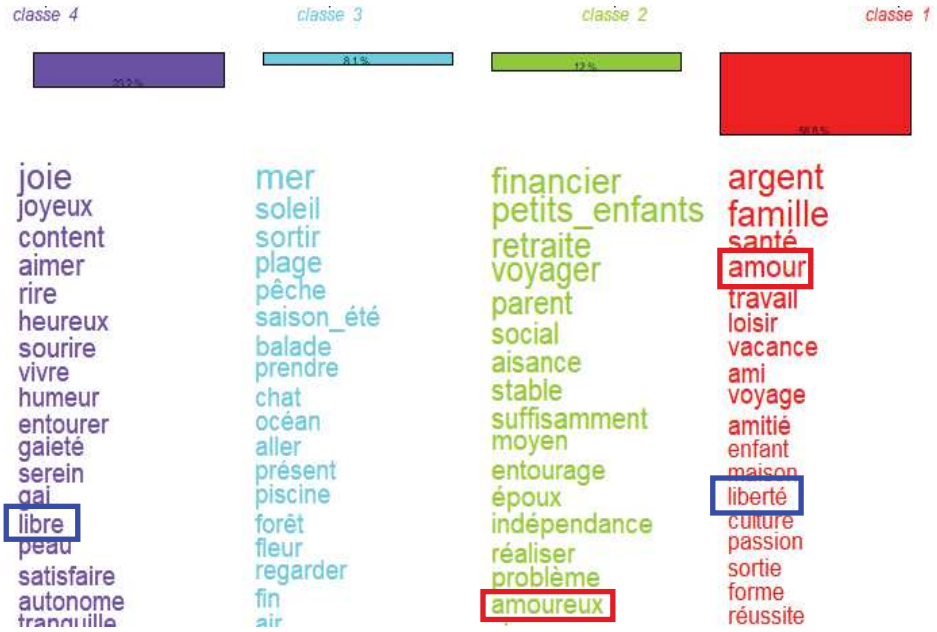

Figure 7 : Figure du logiciel IRaMuTeQ après une CDH sur des données mal lemmatisées (IRaMuTeQ)

En plus de cela, si un mot n'est cité qu'une seule fois par un unique individu, cet individu est « mis à part » et considéré comme un individu rare. Or, avec la méthode de projection sur le premier axe factoriel, l'individu rare va être la proie de la coupure et créer une classe à lui tout seul. Sur le jeu de données de 2013, si l'on garde les mots cités une seule fois, le graphique suivant est affiché (Figure 8a). Il oppose le mot « gai » qui n'est cité que par l'individu 23209 et le mot « insouciant » qui n'est cité que par l'individu 33529. La suite de la procédure consiste à projeter tous les individus sur l'axe 1, et de déterminer la partition optimale en divisant le jeu de données en 2 groupes d'individus par une maximisation du chi2. Pour chaque partition, le chi2 est calculé et retenu afin d'obtenir la courbe de l'évolution du chi2 (Figure 8b).

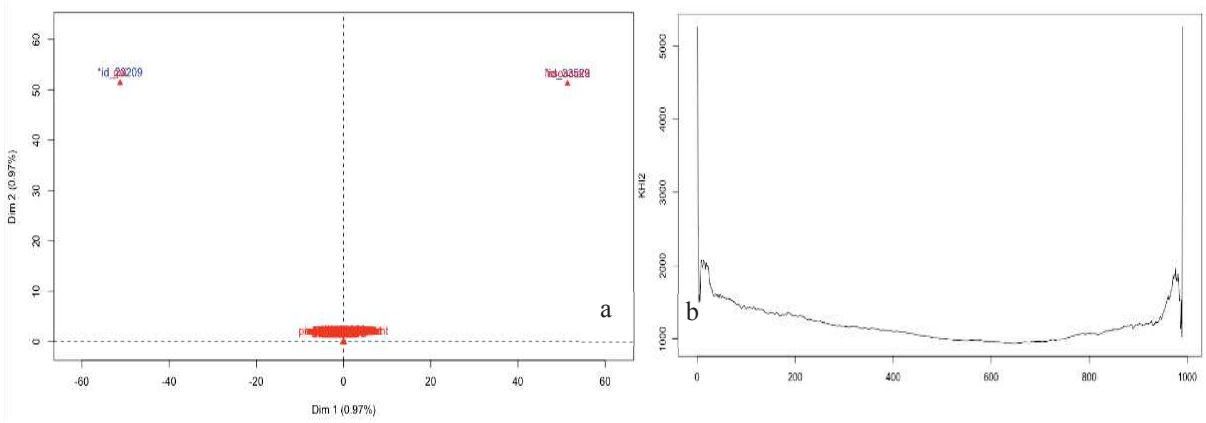

Figure 8 : AFC du jeu de données 2013 lemmatisé avec une fréquence minimum de 1 (a) et l'évolution du chi2 selon la coupure sur le premier axe factoriel (b)

Les deux valeurs maximales du chi2 se trouvent aux deux extrémités du jeu de données. Cela impose donc de couper soit entre le premier individu et tout le reste, soit d'exclure le dernier individu, ce qui a pour conséquence la création d'une classe d'un seul individu, ce qui n'est pas le but ici. Pour éviter ce problème, on supprime les mots ayant une fréquence inférieure à 3 (TJS) du jeu de donnée (Reinert, 1990) (BELDAME et al., 2014).

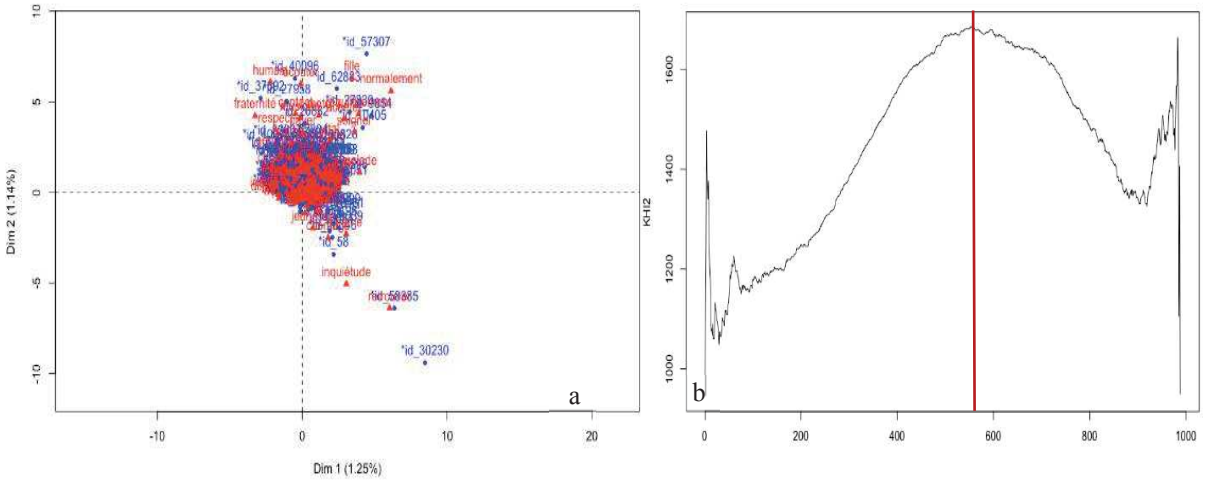

Figure 9 : AFC des lemmes cités au moins 3 fois (a) et évolution du chi2 selon la coupure (b)

Lorsque l'on impose la fréquence minimum d'une forme analysée supérieure strictement à 2, la séparation en deux groupes est plus optimale pour la suite (Figure 9).

C) COMPREHENSION DES CRITERES DE IRAMUTEQ

Ci-dessous (Figure 10) une capture d'écran des paramètres du logiciel d'IRaMuTeQ. L'unique classification que j'effectuerai sera « simple sur texte ». Cette classification permet de diviser les réponses en plusieurs classes lexicales tout en imposant qu'un individu n'appartienne qu'à une unique classe. On peut donc par cette classification faire à la fois une typologie des discours et une typologie de population.

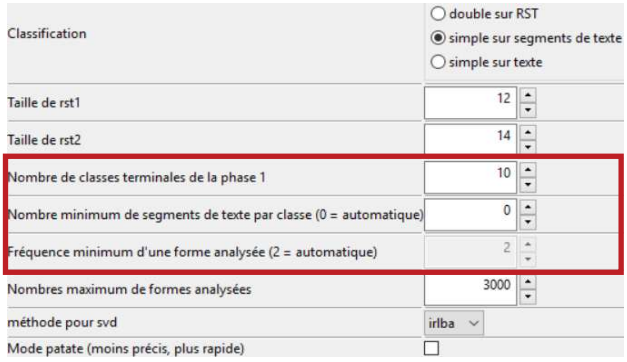

Figure 10 : Fenêtre de paramétrage de la classification descendante hiérarchique (IRaMuTeQ)

Les paramètres utilisables pour cette classification sont mis en évidence en rouge (Figure 10) :

- Nombre de classes terminales : indique le nombre de classes finales de la classification
- Nombre minimum de segments de texte par classe : détermine un seuil minimal de segments de texte en dessous duquel les classes ne seront pas retenues
- Fréquence minimum d'une forme analysée (TJS), fixée supérieure strictement à 2 par Max Reinert (1983)

1) Nombre minimum de segments de texte par classe – un critère ajustable

Parfois, certains mots sont cités plus de 2 fois mais la partition en 2 crée deux classes très peu équitables (10 individus contre 400 par exemple). La classe ne possédant que 10 individus sur 1000 ne représente qu'un pour cent de la population, ce qui, pour les études au sein du CRÉDOC, ne semble pas intéressant (un chi2 entre une variable à 1% contre 40% ne peut pas être significatif car la classe à 1% est trop petite pour être expliquée). Il n'est pas fiable d'essayer de décrire une classe de 10 individus Il est d'usage de caractériser une classe d'individus lorsque le nombre d'individus qui la composent est suffisant (n > 50 par exemple ou 50 pour mille individus interrogés = 5%). Pour cela, le paramètre « Nombre de segments de textes par classe » existe et est un critère ajustable. Si on laisse le paramètre sur « automatique » (0), on obtient un nombre minimal de segments de texte à analyser trop grands. Pour 1000 individus divisés en 5 classes il force chaque classe à avoir au moins 200 individus, or il n'est pas nécessaire d'avoir d'autant pour caractériser une classe. De plus, cela pourrait masquer certaines représentations de l'objet. Si l'on veut au minimum 5% de la population dans une classe, et que l'on a 1000 individus, on fixera ce paramètre à 50. Il forcera la typologie à ne créer que des classes de plus de 50 individus ayant cité des mots qui sont évoqués au moins 3 fois par l'ensemble des individus.

2) Vérification du critère « fréquence minimum d'une forme analysée »

Grâce au code que j'ai développé pour recréer la CDH, j'ai voulu vérifier le critère « fréquence minimum de formes analysées » ou TJS. En effet, selon Max Reinert (1983), il y a une bonne stabilité des résultats de la typologie à partir de TSJ valant 3, soit la fréquence minimum de formes analysées est supérieure ou égale à 3 (Figure 11).

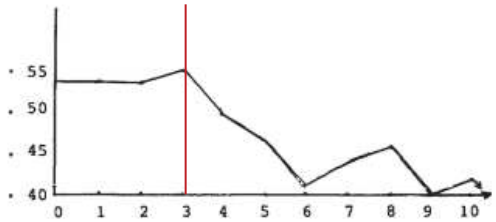

Figure 11 : Evolution de l'inertie en augmentant la fréquence minimum d'une forme (Reinert, 1983)

Pour vérifier sa théorie, j'ai fait fonctionner mon programme en retirant à chaque fois les mots cités moins souvent qu'une fréquence imposée allant de 1 à 10. La première fois que le programme est lancé, tous les mots sont pris en compte dans l'analyse, puis tous les mots cités au moins deux fois, jusqu'à au moins 10 fois. On ne récupère que 4 classes à chaque fois. On calcule la valeur du chi2 sur le tableau croisé « classes x lemmes ». On obtient la courbe bleue ci-dessous (Figure 12 a et b), elle représente le chi2 de l'indépendance entre les 4 classes obtenues finalement, qu'on nommera « chi2inter », et donc l'indicateur de l'inertie interclasse. On remarque la stabilité des classes pour une fréquence minimum supérieure ou égale à 3.

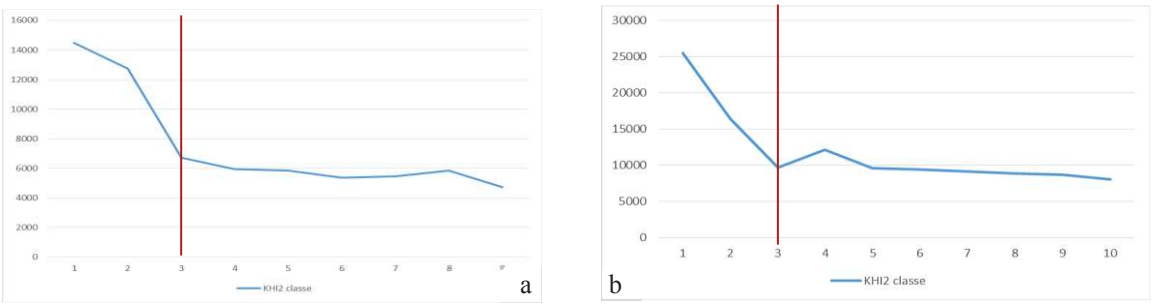

Figure 12 : Evolution du chi2 de l'indépendance sur les 4 classes finales en fonction de la fréquence minimum de formes analysées en 2019 (a) et en 2013 (b)

Je crée un indicateur du pourcentage d'inertie :

$$\frac{\chi^2\ inter}{\chi^2\ total} \sim \frac{Inertie\ interclasse}{Inertie\ totale}$$

« chi2total » représente le chi2 sur le jeu de données initial, équivalent à l'inertie totale. On obtient la courbe de pourcentage d'inertie suivante (Figure 13 courbe jaune).

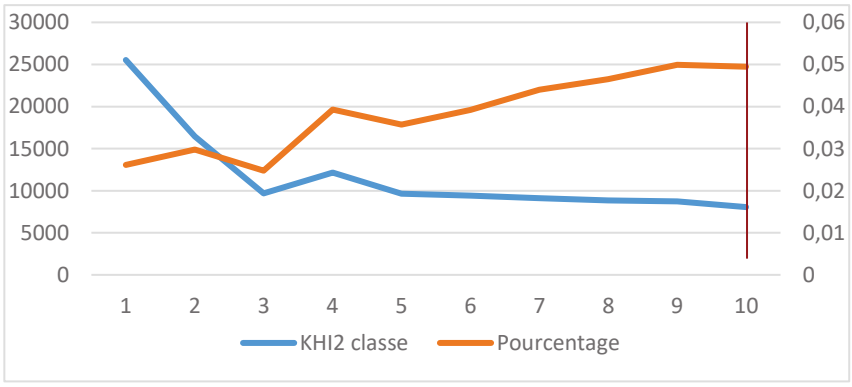

Figure 13 : Evolution du chi2 (courbe bleue) en fonction de la fréquence minimum de formes analysées et du pourcentage chi2inter/chi2total (courbe orange)

Comme l'on souhaite maximiser l'inertie interclasse, il nous faut un chi2 maximal et un pourcentage d'inertie maximal. Il serait donc préférable, ici, de prendre TJS supérieur à 10. Or, si l'on regarde en détail les classes données par l'algorithme, on se rend compte que le nombre de lemmes analysés, pour cette valeur de TJS, est inférieur à 10% du nombre total de lemmes, ce qui est loin d'être optimal. Il faudrait donc, en terme de poids, mettre une importance sur le pourcentage de lemmes analysés (Figure 14a). A la vue de ce nouveau graphique, j'opterai pour une fréquence de formes analysées supérieure ou égale à 1 ou 2.

A nouveau, cela ne convient pas car pour 3 classes parmi les quatre, celles-ci possèdent moins de 1% du jeu de données. Il est important de le prendre en considération car je souhaite des classes ayant au moins 5% de la population afin de pouvoir les décrire ensuite. On part d'un poids de 1. Pour chaque classe possédant moins de 5% de la population étudiée, on ajoute 1. Si 3 classes sur 4 possèdent 1 individu, le poids de la pénalité est de 4 (Figure 14b).

La pénalité est imposée de la manière suivante (Figure 14 a et b):

- Pénalité sur le pourcentage de lemmes analysés (Figure 14a) :

$$\frac{\chi^2 inter \times pourcentage\ de\ lemmes\ analysés}{\chi^2 total}$$

- Pénalité sur le nombre d'individus par classe (Figure 14b) :

$$\frac{\chi^2 inter}{\chi^2 total \times poids\ des\ classes}$$

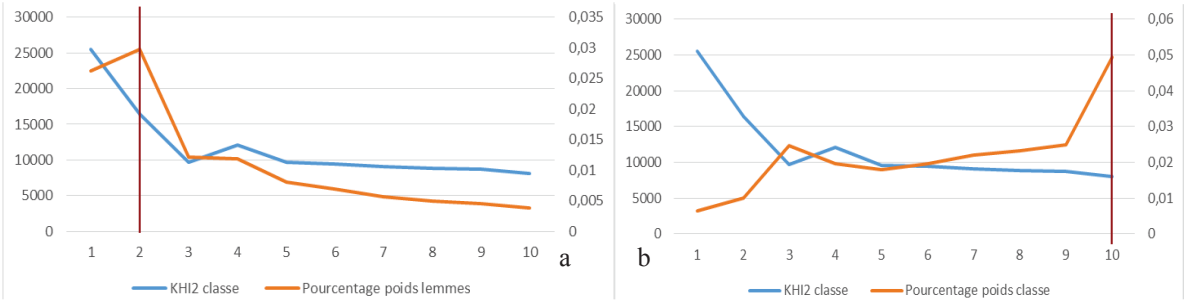

Figure 14 : Evolution du chi2 et du pourcentage de chi2inter pondéré par une pénalité sur les lemmes (a), par le nombre de personnes dans chaque classe (b) en fonction de la fréquence minimum

En prenant en compte tous les aspects du problème, nous considérons le nombre de lemmes plus important que le nombre de personnes par classe car il montre la richesse du vocabulaire et des mondes lexicaux. Pour cela, nous pénalisons (arbitrairement) le pourcentage d'inertie par le carré du pourcentage de lemmes analysés. Ainsi je décide de diviser le pourcentage d'inertie par le nombre de classe + 1 n'ayant pas au moins 5% des individus de la population, et de le multiplier par le pourcentage de lemmes analysés au carré (Figure 15).

Soit : $$\frac{\chi^2 inter \times (pourcentage\ de\ lemmes\ analysés)^2}{\chi^2 total \times poids\ des\ classe}$$

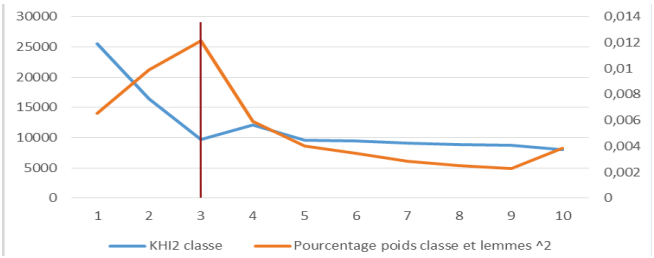

Figure 15 : Evolution du chi2 et du pourcentage chi2inter pondéré par plusieurs pénalités en fonction de la fréquence minimum

Je retrouve bien l'hypothèse que Max Reinert faisait déjà en 1983 (Figure 11). Les classes sont plus stables si la fréquence minimale des formes analysées vaut 3.

3) Recommandations pour le CRÉDOC

Au vu de ce que nous avons pu observer concernant les différents paramètres, il y a certains points sur lesquels le CRÉDOC doit être vigilant lors de ses enquêtes. Il est primordial d'inciter le répondant à citer au moins 5 mots, car si celui-ci ne cite qu'un mot et qu'il est peu cité par les autres individus, il va créer une classe à lui seul ou se retrouver « non-classé ». Il est important de faire attention à la lemmatisation d'IRaMuTeQ. En effet ce logiciel différencie les mots selon leur classe grammaticale. Il faut donc forcer la lemmatisation par modification du dictionnaire. Plus les mots sont lemmatisés et moins les individus rares et hapax sont présents au moment de l'analyse. Il est important de bien connaître les critères de la CDH et son fonctionnement afin de générer la classification la plus stable et la plus intéressante possible.

III. EVOLUTION DU DISCOURS DE 1993 A 2019

Le CRÉDOC utilise, pour comprendre les représentations sociales associées à l'alimentation, plusieurs dispositifs d'enquêtes dont l'enquête CCAF (Comportements et Consommations Alimentaires des Français). Un questionnaire « Tendances de consommation » sonde les comportements alimentaires des individus. La première question posée aux individus de l'enquête de cette année est « *Si je vous dis "être heureux", à quoi pensez-vous ?* ». Les réponses à cette question constituent le matériel textuel support des travaux réalisés au cours de ce stage.

A) DESCRIPTION DES DONNEES

Je dispose de différents jeux de données. J'ai les réponses à l'enquête de consommation de 2013 et celles de 2019. Pour chacune, elle débute par la question ouverte : *« Si je vous dis être heureux à quoi pensez-vous ? »*, utilisée pour la première fois en 1993. Chaque enquête possède ensuite un ensemble de questions sur la condition de vie du répondant et sur ses modes de consommation. En 2013, l'enquête a été effectuée par téléphone, 1012 individus remplissant les quotas imposés par le CRÉDOC ont répondu. En 2019, l'enquête a été faite en ligne (panel), 1497 individus ont répondu tout en remplissant les critères de quotas (sondage par la méthode des quotas).

B) LES RESULTATS DE 1993, 2013 ET 2019

Avant toute étude de l'évolution, il est important d'étudier les résultats de chaque année. Ceci se fait en lisant les cahiers de recherches de 1993 (Lahlou et al., 1993) et 2013 (SIOUNANDAN et al., 2013) et en effectuant les analyses de 2019.

1) En 1993

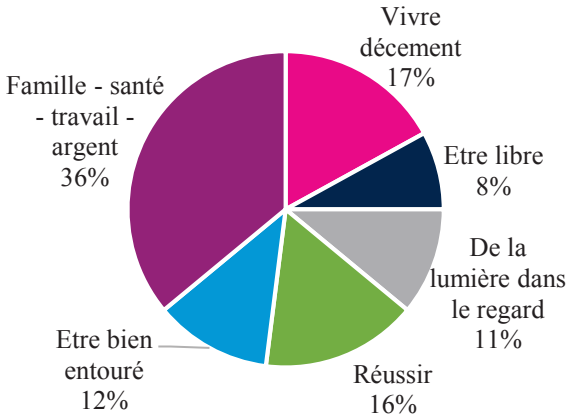

Figure 16 : Typologie de discours de 1993

En 1993, les mots les plus cités sont « famille », « santé », « argent », « travail », « enfants », « amour », « problème », et « réussir ». Après analyse, 6 classes de typologie de discours apparaissent (Figure 16). La première classe (17% de la population interrogée) a été intitulée « **vivre décemment** ». Les individus qui la composent considèrent que le bonheur consiste à ne manquer de rien. La seconde classe, la plus conséquente (36%), parle de « **famille, santé, travail et argent** ». Elle considère le bonheur selon quatre piliers de la culture humaine. La troisième classe se veut « **bien entourée** » (12%), elle a besoin d'affection. Ce sont majoritairement des personnes âgées qui souhaitent une persistance dans le réseau familial afin d'assurer leur sécurité logistique et affective. La quatrième classe (16%) voit son bonheur à travers la « **réussite** ». Le bonheur de ces personnes se calcule en fonction de leur réussite familiale et professionnelle. La cinquième classe (11%) a été nommée « **de la lumière dans les yeux** » et regroupe deux façons de représenter le bonheur : l'une consistant à s'échapper du monde stressant qui nous entoure, l'autre, se voulant pleine de sérénité et de calme. La dernière classe (8%), se veut « **libre** ». Le bonheur en 1993 se mesurait principalement sur le bien-être matériel 1993 (Lahlou et al., 1993).

2) En 2013

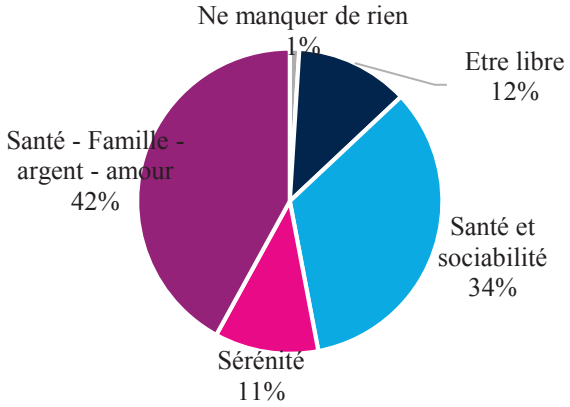

Figure 17 : Typologie de discours de 2013

En 2013, les mots les plus employés sont « santé », « famille », « argent », « enfants », « travail », « loisirs », « vacances », « voyage », « liberté », « petits-enfants », « soucis », « plaisir », « épanouir » et « tranquillité ». Lors de la typologie de discours, 5 classes apparaissent (Figure 17). On retrouve en classe prédominante (42% de la population) une classe autour des grands piliers de la culture humaine : « **santé, famille, argent, amour** ». On retrouve aussi une classe « **être libre** » (12%) : pour eux, le bonheur consiste à outrepasser les contraintes telles que le temps et l'espace. Cette classe est majoritairement constituée de personnes âgées. Une classe « **santé et sociabilité** » est constituée de 34% de la population. Les plus représentés dans cette classe sont les plus âgés, ces personnes trouvent leur bonheur dans la bonne santé et leur entourage. La quatrième classe (11%) se veut « **sereine** ». Ce groupe n'a pas besoin matériel ou social pour être heureux, les individus le composant aspirent à une finalité psychologique. Les personnes de cette classe ont un revenu élevé et possèdent une situation matérielle convenable. La dernière classe « **ne veut manquer de rien** » (1%) est majoritairement représentée par des personnes n'ayant pas de forts revenus, et n'aspirent qu'à mener une vie où tous leurs besoins sont assouvis (SIOUNANDAN et al., 2013).

3) En 2019

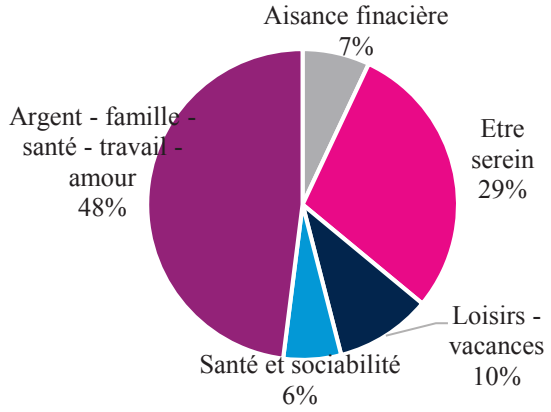

Figure 18 : Typologie de discours de 2019

En 2019, 19% de la population étudiée se dit peu heureuse voire pas du tout heureuse au contraire de 16% d'autres personnes qui se trouvent très heureuses. Les lemmes majoritairement cités sont équivalents à ceux de 2013. On observe aussi une typologie du discours en 5 classes (Figure 18). La première classe est une classe déjà présente en 1993 et en 2013 : « **Argent, famille, santé, travail** », qui comprend 48% de la population. La seconde classe (29%) est une classe qui contient des personnes aisées qui cherchent à atteindre un état de « **sérénité** » ou alors, qui contient des personnes ayant des revenus faibles et qui cherchent à ne plus avoir de soucis. On tourne principalement autour des concepts de liberté et d'indépendance. La troisième classe (10%) est une classe qui trouve son bonheur dans les « **loisirs et les vacances** ». Elle apprécie avoir du temps libre. La quatrième classe (7%) est une classe qui n'est pas très heureuse et qui s'inquiète de son « **aisance financière** ». L'argent est la clé de ses tracas. La dernière classe (6%) est une classe où les plus âgés prédominent et où le bonheur passe par l'« **entourage familial** » et la vie à la campagne.

Il est à ce point important de savoir que le vieillissement *stricto sensu* d'une typologie est impossible (FOURNIER, 2009). Il faut donc trouver un moyen de comparer les périodes et voir l'évolution des discours tout en ayant une méthode applicable.

C) L'EVOLUTION ENTRE 2013 ET 2019

Pour entrevoir l'évolution des mondes lexicaux entre les années 2013 et 2019, il est possible de regarder l'évolution de l'effectif de chaque lemme, c'est-à-dire, le nombre de fois que celui-ci a été cité sur l'année mais aussi, l'apparition et la disparition de certains mots.

La première étape de notre analyse consiste à compter le nombre de lemmes différents, puis, de comparer leurs distributions entre 2013 et 2019. Comme le nombre d'individus total varie en fonction des années, il est nécessaire de trouver un moyen de les comparer. Je relève donc le rang qui classe les lemmes par ordre décroissant de citation (Tableau 6). Il suffit ensuite de comparer les rangs du même lemme entre les deux années, et d'observer les différences. Il est aussi possible de calculer l'effectif relatif de chaque lemme pour comparer sa distribution.

Tableau 6 : Comparaison des classements des termes employés entre 2013 et 2019

	2019 (1497 individus)			2013 (1012 individus)			Evolution 2013-2019	
	Rang	Effectif	Effectif relatif	Rang	Effectif	Effectif relatif		
Famille	1	642	43%	2	374	37%	1	↗
Argent	2	535	36%	3	295	29%	1	↗
Santé	3	485	32%	1	420	42%	-2	↘
Amour	4	386	26%	9	136	13%	5	↗
Ami	5	364	24%	7	139	14%	2	↗
Enfant	6	312	21%	4	219	22%	-2	↘
Bonheur	**7**	**310**	21%	**14**	**82**	8%	**7**	↗
Travailler	8	238	16%	5	204	20%	-3	↘
Vacance	9	234	16%	8	138	14%	-1	↘
Joie	**10**	**228**	15%	**27**	**35**	3%	**17**	↗

Comparativement, les lemmes tournant autour de la « famille », l'« argent », la « santé », l'« amour », les « enfants » et le « travail » restent stables. Les termes liés au « bonheur » sont en pleine explosion, tout comme ceux corrélés au « bien-être ». Les termes « nature », « sport », « animaux » et « sexe » sont beaucoup plus présents. Un dernier type de lemme est plus souvent cité en 2019 ; il s'agit des lemmes évoquant les vacances et les loisirs.

A l'inverse, les termes « couple », « souci », « problème », « réussir », « profiter », « temps », « confort », « social », « stabilité » et « projet », ont chuté dans les rangs.

Beaucoup de mots apparaissent entre 2013 et 2019. Les termes « sain » et « nourriture », « air pur », « forêt », « verdure », « randonnée », « vélo » mais aussi « polluer », rendent compte de l'évolution des préoccupations de la population française (Tableau 7).

Tableau 7 : Apparition et disparition de mots entre 2013 et 2019

Cité en 2019	Rang	Effectif
Sain	77	14
Nourriture	78	14
Chance	97	12
Été	104	10
Zen	112	9

Cité en 2013	Rang	Effectif
Temps-libre	32	30
Entendre	102	7
Épargner	105	6
Trouver	106	6
Changer	119	6

Pour aller plus loin, je peux à l'aide d'IRaMuTeQ et d'un test du chi2, mettre en évidence les mots significativement les plus cités en 2013 et en 2019 (Figure 19). Les lemmes « loisir », « temps-libre », « avenir », « souci » etc. sont majoritairement plus cités en 2013, à l'inverse des lemmes « joie », « bonheur », « amour », « ami », « bien-être », etc. qui sont plus cités en 2019.

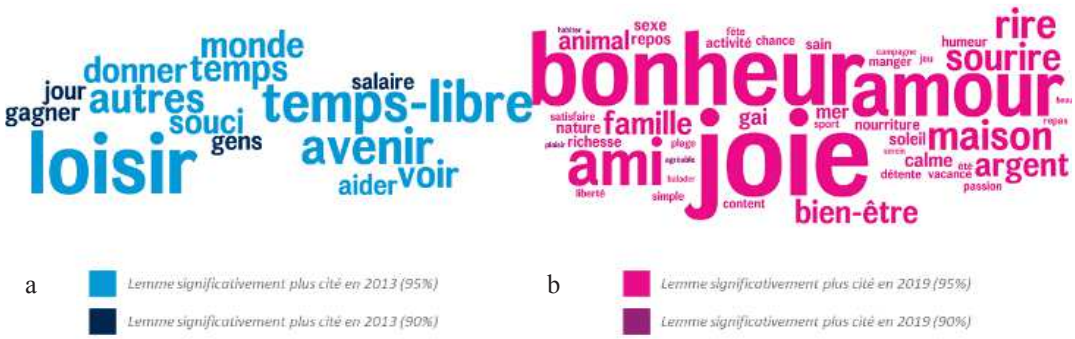

Figure 19 : Mots significativement plus cités en 2013 (a) et en 2019 (b)

Ensuite, j'ai voulu observer l'évolution des mots cités entre les années mais aussi en fonction des variables sociodémographiques (PILORIN et al., 2008). Un simple comptage des lemmes m'indique l'évolution des concepts utilisés pour « être heureux ». En effet, étudier toute la réponse de l'individu dans le sens syntaxique n'apporte pas grand-chose de plus sur ces concepts. Aussi ai-je dû réfléchir à un moyen pour faire apparaître sur un même graphique les lemmes de 2013 et de 2019 ainsi qu'une distinction de ces lemmes selon des variables sociodémographiques. J'ai alors pensé à faire une AFC, à l'aide du package FactoMineR de R, sur un tableau de contingence croisant les lemmes en individus, et, en colonnes, les modalités de la variable sociodémographique souhaitée en différenciant 2013 et 2019. Par exemple, compter pour chaque lemme combien de fois il a été cité par les femmes ou par les hommes, pris séparément, en 2013 comme en 2019. A cela, on ajoute deux colonnes, mises en variables supplémentaires de l'AFC : l'effectif total de chaque lemme par année, afin d'avoir le point moyen de chacune des années (Tableau 8). Tout cela est réalisé à l'aide de la fonction textual du package FactoMineR de R et de la fonction merge.

Tableau 8 : Exemple d'un tableau de contingence soumis à l'AFC pour étudier l'évolution du vocabulaire

	Variables actives				*Variables supplémentaires*	
	Homme2013	Femme_2013	Homme_2019	Femme_2019	*2013*	*2019*
Famille	209	165	374	361	*281*	*642*
Loisir	71	82	153	49	*51*	*100*

Après avoir vérifié que les populations étaient homogènes pour chaque année et pour chacune des variables sociodémographiques, je peux effectuer mes analyses. Je décide de retirer les lemmes cités moins de 15 fois pour n'avoir accès qu'aux grands concepts « d'être heureux ». Plusieurs essais ont été réalisés sur diverses variables sociodémographiques ; les graphiques que je présente ici (Figure 20 et Figure 21) s'avèrent être les plus intéressants.

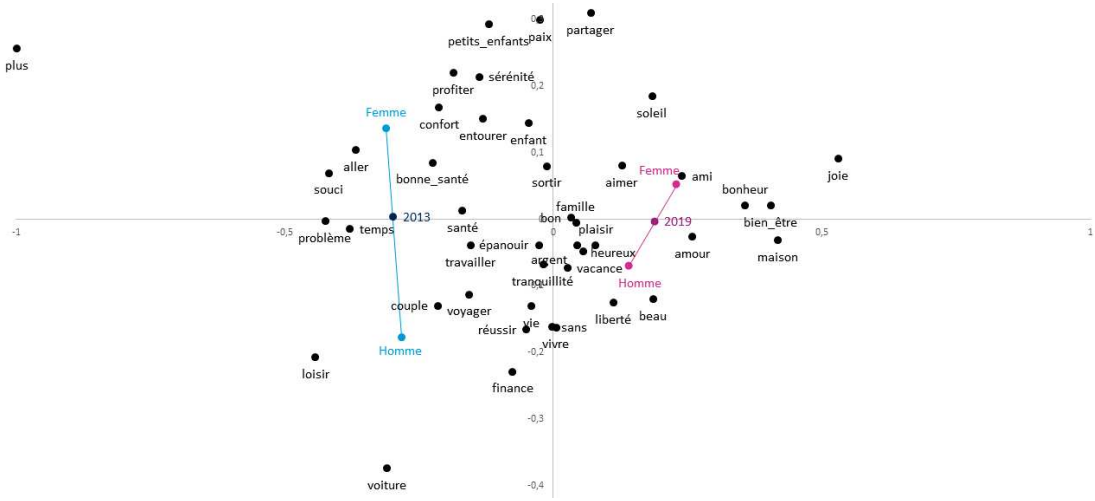

Figure 20: Graphique de l'AFC croisant les lemmes et la variable sexe entre les années 2013 et 2019

Ce graphique (Figure 20) montre que la principale dimension de variabilité est l'année d'étude. La deuxième, sépare les groupes en fonction du sexe. On peut remarquer que l'évolution du vocabulaire des femmes a été plus forte que l'évolution de celle des hommes (points plus éloignés). On peut aussi observer que la différence de vocabulaire entre les deux sexes en 2013 est bien plus prononcée qu'en 2019. En 2013, on se considère être heureux si l'on n'a pas de problème particulier et qu'on est en bonne santé, alors qu'en 2019, le bonheur est un état de bien-être, entouré de ceux qu'on aime. Les hommes ont tendance à penser le bonheur comme une réussite financière et amoureuse, ainsi que de se sentir libre, alors que les femmes veulent profiter de leurs enfants, se sentir en paix, être sereines et surtout entourées.

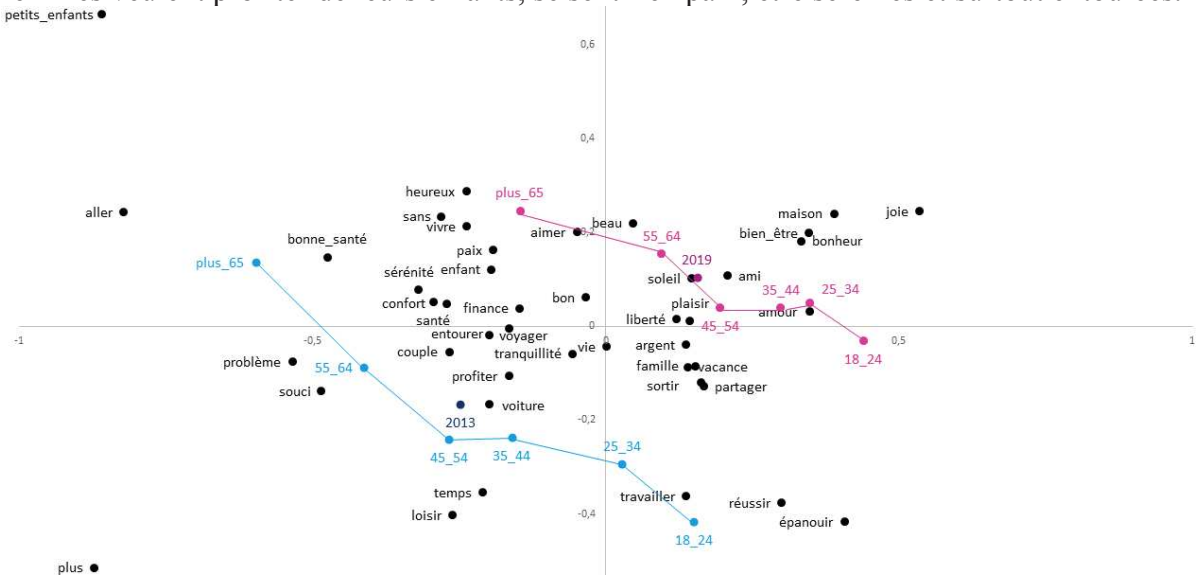

Figure 21 : Graphique de l'AFC croisant les lemmes et la variable âge entre les années 2013 et 2019

Pour ce qui est de l'âge (Figure 21), les différences se situent sur les diagonales des axes de l'analyse factorielle. La première diagonale oppose 2013 (en bas à gauche) et 2019 (en haut à droite). La deuxième diagonale oppose les plus âgés (en haut à gauche) et les plus jeunes (en bas à droite). On retrouve l'opposition entre les deux années. En effet, en 2013, les concepts de ne pas avoir de soucis particuliers, être en bonne santé, avoir du temps libre pour les loisirs, travailler et réussir sont prioritaires, tandis qu'en 2019, le bonheur passe par le bien-être, l'amour et l'amitié, la liberté et le plaisir. Sur les deux années, les plus âgées considèrent le bonheur à travers leurs petits-enfants. En 2019, cela passe aussi par le fait de vivre en paix alors qu'en 2013, elles se concentrent plus sur leur bonne santé. Pour les plus jeunes, en 2013, le bonheur se focalise sur le travail, la réussite et l'épanouissement, tandis que pour les jeunes de 2019, il passe par l'amour et le partage. Les différences de vocabulaires entre les différentes classes d'âge sont bien plus prononcées en 2013 qu'en 2019, en atteste les distances entre chaque point. Tout comme en 2013, les plus âgés, en 2019, se singularisent encore du reste de la population. En revanche, on observe un très net rapprochement des classes d'âge 25-34 et 35-44 entre les deux années, tout comme, mais de moindre façon, un rapprochement global de la population, qui tend à penser que la représentation du bonheur s'uniformise entre générations.

Les représentations du bonheur varient peu entre 1993 et 2019. L'amour, l'argent, la famille, les amis et le travail forgent les grands concepts du bonheur. Le changement existant se fait au niveau de la réussite professionnelle. Il n'est plus aussi important de réussir professionnellement pour être heureux. Même si pour certains, les moins aisés, le bonheur reste orienté sur la préoccupation financière de leur foyer, les Français dans l'ensemble désirent davantage voyager, être en vacances et avoir des loisirs pour être heureux. Les notions de liberté et de sérénité sont des valeurs qui ne cessent de prendre de l'ampleur. Étant donné la vie professionnelle stressante et imposante, il est important de trouver un moyen d'évasion pour se sentir heureux. On veut profiter de chez soi, de son entourage et être bien dans sa tête.

DISCUSSION ET CONCLUSION

Plusieurs caractéristiques sont à prendre compte lorsqu'on effectue une analyse lexicale avec le logiciel IRaMuTeQ. Certaines limites se trouvent au niveau de :

- **LA QUALITÉ DU CORPUS.** La différence de récupération des données de la question ouverte pourrait impliquer un biais dans la façon de répondre des enquêtés. Il est apparu que le mot « sexe » est un mot très cité en 2019 (enquête online) alors qu'il est quasiment inexistant en 2013 (enquête téléphonique). Il est difficile de savoir si ce mot est plus cité car la société n'a plus de tabou associé à ce mot ou car l'ordinateur n'est pas un interlocuteur direct et que l'on est plus à l'aise pour en parler. De plus, lors des enquêtes onlines, personne ne peut affirmer l'authenticité des réponses de l'enquêté. Au téléphone, la personne chargée de poser les questions peut décider d'invalider les réponses d'un individu si elles sont incohérentes. L'enquête online donne lieu à des réponses inexploitables telles que des suites de lettres insignifiantes, écrites pour passer à la suite du questionnaire et toucher la récompense associée à son remplissage. Cela représente, en 2019, 4% des réponses qui se retrouvent automatiquement non classées.
- **LA LEMMATISATION.** Plusieurs limites apparaissent ici. Pour les enquêtes effectuées au CRÉDOC, la lemmatisation distinguant les classes grammaticales n'est pas optimale. Il faut donc à la main, modifier le dictionnaire de lemmatisation afin de ne plus différencier les mots en fonction de leurs classes. D'autre part, comme vu précédemment avec le cas de la saison « été », il est difficile de savoir quel choix arbitraire effectuer pour laisser apparaître tous les concepts du bonheur des français. Ces décisions sont à prendre de manière méticuleuse grâce aux sociologues de l'entreprise.
- **LES PARAMÈTRES D'IRAMUTEQ.** Le critère de fréquence minimum de forme analysée a lui aussi ses limites. Ne pas analyser les lemmes de fréquence inférieure à ce critère pourrait faire disparaître des représentations du bonheur. Il pourrait être profitable de voir comment pallier ce problème. J'aurais pu essayer de dupliquer 3 fois le jeu de données et analyser la variation des classes. Il pourrait aussi être intéressant de laisser ce critère libre dans IRaMuTeQ afin de voir, si, pour un corpus donné, il est possible de diminuer cette fréquence. Un des critères utilisés dans la CDH n'est pas non plus ajustable dans IRaMuTeQ. Il s'agit du coefficient C qui caractérise les classes chevauchantes et dont la valeur optimale est 0,3 (CTEST). J'aurai pu essayer de modifier cette valeur afin d'observer les changements provoqués.
- **LES RESULTATS.** L'évolution d'analyses lexicales portant sur le même sujet (même question entre plusieurs périodes) est compliquée car malgré une forte ressemblance entre les classes, il y aura quasiment à chaque fois une légère modification des modalités qui caractérisent ces classes (FOURNIER, 2009).

Il y a néanmoins des avantages pour le CRÉDOC de passer de ALCESTE à IRaMuTeQ. Ce dernier est tout d'abord gratuit. Avec l'augmentation du volume des corpus, il analyse des corpus de plus grosse taille et il est bien plus rapide qu'ALCESTE (Ratinaud, Marchand, 2012).

Ce rapport étudie les objectifs de l'analyse lexicale et le fonctionnement d'IRaMuTeQ avec en particulier le processus associé à la CDH et ses paramètres. Il a aussi été entreprit d'observer l'évolution du vocabulaire entre deux périodes à l'aide d'AFC sur les lemmes cités. Pour approfondir l'analyse, il aurait pu être intéressant d'étudier cette évolution à l'aide de segments de textes afin d'affiner l'analyse sociologique des représentations du bonheur.

Dans la suite de ma mission, j'effectue à l'aide d'une AFDM et une CAH, une typologie de tendances de consommations qui va me permettre d'observer l'association des représentations du bonheur aux tendances de consommations de 1993 à 2019.

BIBLIOGRAPHIE

BEAUDOUIN, Valérie, 2016. Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données. In : JADT 2016 [en ligne]. Nice, France : Mayaffre, D. Poudat, C., Vanni, L. et al. juin 2016. p. 17-27. [Consulté le 8 août 2019]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01376938.

BELDAME, Diane, HEBEL, Pascale et MATHE, Thierry, 2014. Evolution des représentations sociales du bien manger. In : [en ligne]. 2014. [Consulté le 6 août 2019]. Disponible à l'adresse : https://www.credoc.fr/publications/evolution-des-representations-sociales-du-bien-manger.

BENZECRI, Jean-Paul, 1973. L'analyse des données 2 : L'analyse des correspondances. Dunod. Paris : Dunod. Pratique de l'analyse des données : linguistique et lexicologie. Paris : Dunod, 1981.

FOURNIER, Olivier, 2009. Comparaison de partitions issues de données de comportements alimentaires. S.l.

LAHLOU, Saadi, 1992. Si/alors : « bien manger » ? Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2 000 individus. In : [en ligne]. 1992. [Consulté le 20 juin 2019]. Disponible à l'adresse : https://www.credoc.fr/publications/sialors-bien-manger-application-dune-nouvelle-methode-danalyse-des-representations-sociales-a-un-corpus-constitue-des-associations-libres-de-2-000-individus.

LAHLOU, Saadi, 1993. L'analyse lexicale : - outil d'exploration des représentations. In : 1993. p. 145.

LAHLOU, Saadi, 1993. Réponse à une question ouverte : incidence du mode de questionnement. In : [en ligne]. 1993. [Consulté le 19 juin 2019]. Disponible à l'adresse : https://www.academia.edu/21100726/R%C3%A9ponse_%C3%A0_une_question_ouverte_incidence_du_mode_de_questionnement.

LAHLOU, Saadi, DE BORELY, Aude Collerie et BEAUDOUIN, Valérie, 1993. Où en est la consommation aujourd'hui ? In : 1993. p. 205.

LION, Sébastien et LAHLOU, Saadi, 1991. Construction d'un corpus et perte d'information en analyse lexicale (Méthodes et pratiques). In : [en ligne]. 1991. [Consulté le 20 juin 2019]. Disponible à l'adresse : https://www.credoc.fr/publications/construction-dun-corpus-et-perte-dinformation-en-analyse-lexicale-methodes-et-pratiques.

LOUBÈRE, Lucie et RATINAUD, Pierre, 2014. Documentation IRaMuTeQ 0.6 alpha 3 version 0.1. S.l.

PILORIN, Thomas, HÉBEL, Pascale et MATHE, Thierry, 2008. Du discours nutritionnel aux représentations de l'alimentation. In : [en ligne]. 2008. [Consulté le 6 août 2019]. Disponible à l'adresse : https://www.credoc.fr/publications/du-discours-nutritionnel-aux-representations-de-lalimentation.

RATINAUD, Pierre, 2018. Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. S.l.

RATINAUD, Pierre et MARCHAND, Pascal, 2012. Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRAMUTEQ. In : 2012. p. 10.

REINERT, A, 1983. Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. In : 1983. p. 13.

REINERT, Max, 1987. CLASSIFICATION DESCENDANTE HIERARCHIQUE ET ANALYSE

LEXICALE PAR CONTEXTE - APPLICATION AU CORPUS DES POESIES D'A. RIMBAUD. In : BMS : Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique. 1987. n° 13, p. 53-90. JSTOR

REINERT, Max, 1990. Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. In : 1990. Vol. 15, n° 1, p. 17.

REINERT, Max, 1993. Les « mondes lexicaux » et leur 'logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. In : Langage & société. 1993. Vol. 66, n° 1, p. 5-39. DOI 10.3406/lsoc.1993.2632.

REINERT, Max, 2008. Mondes lexicaux stabilisés et analyse statistique de discours. In : 2008. p. 13.

SESSEGO, Victoire et HEBEL, Pascale, 2018. Consommer durable est-il un acte de distinction ? Représentations, pratiques et impacts écologiques réels au regard des dynamiques sociales. In : [en ligne]. 2018. [Consulté le 6 août 2019]. Disponible à l'adresse : https://www.credoc.fr/publications/consommer-durable-est-il-un-acte-de-distinction-representations-pratiques-et-impacts-ecologiques-reels-au-regard-des-dynamiques-sociales.

SIOUNANDAN, Nicolas, HEBEL, Pascale et COLIN, Justine, 2013. Va-t-on vers une frugalité choisie ? In : [en ligne]. 1 décembre 2013. [Consulté le 6 août 2019]. Disponible à l'adresse : https://www.credoc.fr/publications/va-t-on-vers-une-frugalite-choisie.

YVON, François, 1990. L'ANALYSE LEXICALE APPLIQUEE A DES DONNEES D'ENQUETE ETAT DES LIEUX. In : 1990. p. 151.

SITOGRAPHIE

IRaMuTeQ : http://iramuteq.org/
CRÉDOC : https://www.credoc.fr/

	Diplôme : **Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage** Spécialité : **Agronomie** Spécialisation / option : **Science des données** Enseignant référent : **François HUSSON**

Auteur(s) : **Six Chloé** Date de naissance : **19/03/1996**	Organisme d'accueil : **CRÉDOC** Adresse : **142 rue de Chevaleret** **75013 Paris** Maître de stage : **Gabriel TAVOULARIS**
Nb pages : **20** Annexe(s) : **Aucune** Année de soutenance : **2019**	

Titre français : **Analyse lexicale appliquée à une question ouverte à l'aide d'IRaMuTeQ**

Titre anglais : **Lexical analysis applied to an open-ended question using IRaMuTeQ**

Résumé (1600 caractères maximum) :

IRaMuTeQ est le successeur gratuit d'Alceste, deux logiciels d'analyse lexicale. Il offre un grand nombre d'analyses. Après une lemmatisation du corpus, il propose de compter le nombre de fois que chaque lemme a été cité, de comparer entre les modalités d'une variable « hors corpus » la différence significative de vocabulaire utilisé par chacune par un test du chi2. Il permet aussi de faire une typologie de discours à l'aide de la classification descendante hiérarchique développée par Max Reinert (1983). Cette méthode débute par une AFC, puis continue par la recherche optimale d'une partition pour aboutir à la séparation du corpus en un nombre de classe voulu. Certains paramètres (ajustables ou non) doivent être compris afin d'utiliser correctement le logiciel : le nombre minimum de segments de texte par classe, la fréquence minimum d'une forme analysée et le critère de chevauchement. Ce logiciel a permis d'analyser les données de l'enquête « Tendances des consommations » du CRÉDOC de 2019. Le but étant de comprendre les représentations mentales et sociales du bonheur grâce à la question ouverte « Si je vous dis « être heureux », à quoi pensez-vous ? ». Il a été intéressant d'étudier comment comparer des données sur plusieurs années. La même question ayant été posée en 2013, et sachant qu'une typologie ne peut être vieillie, un système d'AFC appliqué sur un tableau de contingence entre les lemmes et les variables sociodémographiques à étudier par année, permet de voir l'évolution du vocabulaire entre ces deux années.

Abstract (1600 caractères maximum):

IRaMuTeQ is the free successor to Alceste. They are two lexical analytics software. IRaMuTeQ offers many analyses. After a lemmatization of the corpus, it provides to count the number of times each lemma has been quoted. It also offers to compare between the modalities of a variable "*out of corpus*" the significant difference in the vocabulary used by each one by a Chi2 test. It is also possible to make a language typology using the hierarchical top-down classification developed by Max Reinert (1983). This method begins with a CA, then continues with the search for an optimal partition and results in the separation of the corpus into a desired class number. Some parameters (adjustable or not) must be understood in order to use the software correctly: the minimum number of text segments per class, the minimum frequency of an analyzed form, and the overlapping criterion. Thanks to this software, I analyzed data from the Consumer Trends survey developed by the CREDOC. The goal is to understand the mental and social representations of happiness through the open question "*If I tell you "being happy", what are you thinking about?*". The same question has been asked in 2013, and knowing that a typology cannot be aged, a system of AFC applied on a contingency table between lemmas and sociodemographic variables to be studied per year, allows to see the evolution of vocabulary between these two years, with the possibility to see these evolutions according to socio-demographic variables.

Mots-clés : **Analyse lexicale, IRaMuTeQ, Alceste, CRÉDOC, AFC, Classification descendante hiérarchique.**

Key Words: **Lexical analysis, IRaMuTeQ, Alceste, CRÉDOC, CA, Hierarchical top-down classification.**

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*