« Thlemmatiser » un corpus avec Iramuteq : hybrider thèmes et lemmes pour faciliter la lecture des données (le cas des coupes menstruelles)

Franck Cochoy[1]

[1]Université Toulouse Jean Jaurès et IUF – cochoy@univ-tlse2.fr

Abstract 1

Based on a corpus of reviews posted on Amazon.com about a menstrual cup, the paper presents a tip aimed at easing the reading of graphs produced by Iramuteq-assisted similarity analysis, but also to help the software identify thematic units that bring together disjointed series of words. The process consists of identifying the lexical fields that are the primary focus of the search, and substituting these lexical fields, or themes, for the lemmas in the dictionary. The operation significantly reduces the number of forms displayed for a given frequency, and also strengthens the uncovering of links between these thematic units and the associated words.

Keywords: Iramuteq, similarity analysis, themes, lemmas, lexical fields.

Abstract 2 (in French)

La communication présente, à partir de l'exemple d'un corpus d'avis postés sur Amazon.com à propos d'une coupe menstruelle, une astuce destinée à faciliter la lecture des graphiques produits lors d'analyses de similitude par Iramuteq, mais aussi à aider le logiciel à identifier les unités thématiques qui rassemblent des séries de mots disjoints. La démarche consiste à identifier les champs lexicaux sur lesquels porte prioritairement la recherche, et à substituer ces champs lexicaux, ou thèmes, aux lemmes du dictionnaire. L'opération réduit significativement le nombre de formes affichées pour une fréquence donnée, et renforce aussi la mise au jour des liens entre ces unités thématiques et les mots associés.

Mots clés : Iramuteq, analyse de similitude, thèmes, lemmes, champs lexicaux.

1. Introduction

Dans le cadre d'une recherche portant sur les produits d'hygiène jetables (Cochoy, 2021), j'ai analysé un corpus de 5 235 avis de consommatrices postés sur Amazon.com à propos de la Star Cup (pseudonyme), un produit leader sur le marché des coupes menstruelles. Afin de mieux retracer les préoccupations exprimées dans ces avis, j'ai mis en œuvre une approche que je propose d'appeler « thlemmatisation ». Ce néologisme est une contraction de « thématisation » et de « lemmatisation ». L'idée est de coder les formes actives d'un corpus donné selon une liste de catégories thématiques auxquelles elles appartiennent, et de remplacer les lemmes par les thèmes correspondants. Cette approche permet de combiner d'une part une focalisation sur des thèmes de recherche particuliers (par exemple, dans mon cas, les questions de santé, d'économie ou de protection de l'environnement) et d'autre part l'analyse automatique des textes. La thlemmatization n'a en soi rien d'innovant : des fonctionnalités similaires ont déjà été proposés par d'autres logiciels.[1] Il s'agit donc ici d'une contribution modeste, qui vise surtout à rendre compte d'un problème (l'incapacité d'un logiciel à comprendre ce que des mots ont en commun) et à présenter aux utilisateurs d'Iramuteq une astuce permettant sinon de le résoudre, du moins de le pallier à titre partiel.

1. Voir la fonction "EQUIV" de Spad-T au cours des années 1980s disponible dans Spad-T dans les années 1980 et d'autres outils destinés à exécuter des procédures de recodage similaires basées sur des listes proposées par des progiciels plus récents (par exemple, Taltac). Thlemmatiser un corpus est également proche du « topic modelling », c'est-à-dire du marquage des notions clés sur lesquelles se concentre le chercheur, même si le but de la thlemmatisation n'est pas de former un logiciel d'apprentissage automatique à catégoriser des textes qui véhiculent ces notions (ex: avis consommateurs) mais de mieux retracer comment les notions étiquetées sont liées aux autres dans l'ensemble du corpus.

Comme nous le verrons, la « thlemmatisation » d'un corpus est un bon moyen de simplifier ce dernier en fonction des intérêts de recherche *a priori* et d'accroître ainsi la lisibilité et la signification des graphiques obtenus par les procédures textométriques classiques. Après avoir présenté les données et la méthodologie de l'enquête, je procèderai à l'analyse des données choisies et en exposerai les principaux résultats[2].

2. Un corpus d'avis postés sur Amazon

Comment pouvons-nous tenir compte des préoccupations des consommatrices concernant les produits qu'elles consomment ? Dans les développements qui suivent, je propose de me concentrer sur les avis des consommatrices concernant une coupe menstruelle donnée, dont les fabricants vantent aujourd'hui le caractère soi-disant sûr et écologique. Le marché des produits d'hygiène féminine étant un marché mondial, je propose de tracer et d'analyser ces avis au niveau global, en étudiant les évaluations des consommatrices postées sur Amazon.com, étant donné l'importance croissante de ce type d'expression pour les marchés (Beuscart et al., 2016) et la société contemporaine (Blank, 2006). Grâce à une « procédure de web scraping », j'ai récupéré la collection complète des avis de consommatrices postés sur Amazon.com à propos du produit leader sur le marché mondial des coupes menstruelles, soit une collection de 5 235 avis publiés entre 2005 (deux ans après le développement du produit) et 2019. La taille moyenne des commentaires est de 597 caractères, avec un écart type de 739 caractères ce qui est considérable et dénote d'une forte implication d'une partie des autrices. Le nombre de formes est de 6007 dont près de 53 % sont des hapax. Ce taux élevé est dans l'absolu un signe de faible qualité de l'écrit, mais il est conforme à ce que l'on observe généralement dans les textes publiés en ligne par un public large (Loubère et Marchand, 2019).

Les coupes menstruelles appartiennent à une catégorie de produits que les spécialistes du marketing appellent des produits à forte implication (Petty et Cacioppo, 1983), c'est-à-dire des produits dont l'achat et la consommation suscitent une forte réflexivité, des efforts de recherche d'informations, divers essais et expériences, des procédures de calcul et d'évaluation, un investissement en termes de valeurs, et parfois même un engagement moral et politique. Il est donc logique de prendre en compte ces préoccupations, de voir ce qu'elles sont (ou ne sont pas), de regarder la manière dont elles sont exprimées et articulées, et d'évaluer les enjeux associés. Mais il faut se donner les moyens d'une telle analyse, et cette tâche dépasse les capacités d'un seul chercheur, compte tenu de la taille du corpus : la collection complète des avis s'élève à 3,9 millions de caractères, soit 673 633 mots, c'est-à-dire 1 444 pages A4 en simple interligne, courrier 12. Afin de surmonter cette difficulté j'ai donc choisi de recourir à Iramuteq, et de compléter les résultats obtenus avec une analyse de contenu classique.

3. De la lemmatisation à la thlemmatisation

Iramuteq permet de mettre au jour des classes conceptuelles ou des structures relationnelles sous-jacentes à d'énormes bases de données textuelles. L'objectif du logiciel n'est pas de mettre en évidence les mots les plus fréquents, mais de tracer les associations les plus significatives entre eux, quel que soit leur nombre dans le corpus global. Ceci est particulièrement vrai pour l'une des fonctionnalités les plus séduisantes et les plus puissantes d'Iramuteq : la conduite d'une « analyse de similitude ». Ce type de traitement, fondé sur la théorie des graphes, consiste à repérer des communautés de mots fréquemment associés dans

2 Je remercie vivement les deux relecteurs anonymes pour leurs remarques et conseils, ainsi que Pascal Marchand et Pierre Ratinaud pour leur formation, leur assistance, et leur patience. Bien entendu les propos tenus ici n'engagent que moi.

un corpus donné. Une fois que les cooccurrences entre les mots ont été identifiées, le logiciel trace le graphique correspondant. Ce graphique représente le réseau des formes, et met clairement en évidence les sous-communautés de notions fréquemment associées grâce à des couleurs appropriées ou « halo zones » (Marchand et Ratinaud, 2012).

Cependant, lorsqu'on effectue une analyse de similitude sur un corpus de grande taille riche en vocabulaire, on comprend rapidement que le graphique résultant ne sera véritablement lisible que si le nombre de formes prises en compte est réduit ; dans le cas contraire, de nombreuses formes se superposent au risque d'une certaine illisibilité. Afin d'aider le chercheur à restreindre le vocabulaire avant d'effectuer l'analyse, le logiciel propose une liste de formes sélectionnables… triées selon leur fréquence. Bien entendu, il est parfaitement possible de ne pas se concentrer sur la partie supérieure de la liste. On peut cliquer sur les formes disponibles comme on le souhaite, en garder certaines et en ignorer d'autres, mais avec des milliers de formes, une telle sélection manuelle s'avère difficile à mettre en œuvre. On ne peut pas opérer de sélection rigoureuse sans le faire selon un ensemble de critères prédéfinis, ce qui nécessite de connaître la liste complète avant d'effectuer la sélection. Mais, même lorsque ce travail préparatoire a été effectué, la sélection manuelle des mots choisis dans l'ensemble de la liste proposée par Iramuteq s'avère fastidieuse et hasardeuse. Il est important de noter que le logiciel n'est en rien responsable du problème. En effet, étant privé de tout moyen de connaître le sens des mots, Iramuteq ne peut que proposer leur fréquence comme moindre mal pour aider à leur sélection, malgré l'affirmation de la relative non pertinence de ce critère.

Il existe toutefois une autre approche pour surmonter cette difficulté. Au lieu de classer les mots selon leur fréquence, pourquoi ne pas les regrouper selon leur sens ? L'idée est de se concentrer sur les champs lexicaux, et de trouver un moyen d'amener Iramuteq à rendre compte de ces derniers. Cette approche est meilleure que les filtres de fréquence, car elle est purement axée sur le sens. Voici un exemple élémentaire tiré de mon cas. Lors de l'évaluation d'un produit sur un site web commercial, les personnes utilisent d'innombrables qualificatifs, soit positifs (« étonnant », « fantastique », « génial », « formidable », « merveilleux », etc.), soit négatifs (« affreux », « trompeur », « épouvantable », « horrible »[3], etc.) Un logiciel de textométrie ne peut pas savoir ni ce que ces mots ont en commun (c'est-à-dire être des adjectifs de valorisation) ni ce qui les différencie (c'est-à-dire être soit positifs soit négatifs). Pour le logiciel, ces mots ne sont que des mots, comme tous les autres. Bien sûr, étant donné leur signification proche, il est très probable que les adjectifs positifs et négatifs soient associés à des formes similaires dans le corpus, feront partie de structures syntaxiques homologues et apparaîtront donc dans la même zone des graphiques. Mais une probabilité n'est pas une certitude et, dans tous les cas, la superposition graphique de notions proches risque de brouiller inutilement la lecture : des adjectifs similaires se chevaucheront au mieux et seront dispersés au pire, avec le risque de devenir invisibles, alors qu'ils expriment pourtant une même idée, fréquente et forte.

Afin de contrer ces effets et d'aider Iramuteq à prendre en compte le sens des mots, je propose de « thlemmatiser » le corpus auquel ces mots appartiennent. Ce néologisme combine deux notions : les thèmes et les lemmes. Comme nous le savons, un lemme est la racine linguistique commune partagée par un ensemble de formes parentes (« être » est par exemple le lemme de « es », « suis », « sont », « serez », « étions », etc.). Iramuteq est capable de relier les formes aux lemmes correspondants grâce à un tableau sous-jacent (un dictionnaire). De même, compte tenu de son sujet et de ses questions de recherche, un chercheur sait quels mots ont la même signification. Selon les spécialistes d'Iramuteq, « une thématique peut être

3. Dans le corps du texte et pour une meilleure lisibilité, je traduis en français des mots qui figurent en anglais dans le corpus.

définie comme un ensemble de formes pleines cotextuelles liées entre elles par leur objet et leur contexte » (Ratinaud et Marchand, 2015). D'où l'idée : si les thèmes sont importants, pourquoi ne pas, au lieu d'attendre que la procédure de classification n'en identifie que quelques-uns parmi un grand nombre, aider le logiciel à apprendre comment saisir une diversité beaucoup plus grande de significations particulières ? Cet objectif peut être atteint grâce à une reconfiguration du dictionnaire. L'opération consiste à remplacer les lemmes par des thèmes, afin de forcer le logiciel à considérer les différentes formes comme recourant non pas à la même racine (lemme) mais au même champ lexical (thème). De la même manière que la lemmatisation d'un corpus consiste à relier les différentes formes d'un mot donné sous leur racine linguistique, la thlemmatisation d'un corpus consiste à relier les différents mots qui sont utilisés pour exprimer une même idée sous la forme d'un équivalent général. Par exemple, dans l'exemple ci-dessus, un chercheur intéressé par la « sentiment analysis » – c'est-à-dire par la prise en compte des différents sentiments exprimés dans un corpus donné (Liu, 2012) – déclarera le thème « bon » comme le lemme regroupant « étonnant », « fantastique », « génial », « formidable », « merveilleux », etc. Il remplacera donc les lemmes existants du dictionnaire en conséquence. Au cœur de la stratégie de thlemmatisation se trouve un paradoxe intrigant : l'obtention d'une vision plus précise d'un corpus donné (mettre en évidence les préoccupations qui comptent) repose sur une procédure visant à instaurer du flou (fusionner des notions quasi synonymes sous la forme d'un seul équivalent).

Un autre paradoxe est que la procédure n'est réalisable et utile que si elle est appliquée partiellement. En fait, lemmatisation et thlemmatisation vont de pair. D'une part, compte tenu de ses connaissances et objectifs de recherche et de la lecture inductive de l'ensemble du lexique ou du corpus, le chercheur identifie et construit les champs lexicaux qui selon lui méritent d'être retenus comme des thèmes clés sous lesquels une partie du lexique peut être thlemmatisée. D'autre part, le chercheur laisse tous les autres mots inchangés, avec leurs lemmes tels qu'ils existent dans le dictionnaire standard. La thlemmatisation est donc une procédure partielle : une partie du vocabulaire est thlemmatisée ; le reste demeure inchangé (pour les formes restantes on conserve la correspondance entre formes et lemmes du dictionnaire d'origine).

Plusieurs raisons justifient de procéder à une thlemmatisation partielle plutôt qu'à une thlemmatisation complète. Certaines de ces raisons sont triviales : parce qu'un corpus compte des milliers de formes, la définition et le codage des champs lexicaux sont des opérations très chronophages et délicates (il est souvent difficile, voire impossible, de déterminer quel thème pourrait englober certaines notions rares, isolées ou spéciales). Mais ces raisons ne sont pas les principales. Aucun thème n'existe en soi ; contrairement aux lemmes, les thèmes ne sont pas génériques et universels[4] ; leur nombre et leurs définitions dépendent de la recherche en jeu, et sont donc nécessairement limités. De plus, une thlemmatisation partielle permet de mettre en évidence les thèmes choisis dans le corpus. Parce qu'un thème donné rassemble et remplace plusieurs notions sous-jacentes, sa fréquence s'élève à la somme des formes thlemmatisées, et rend ainsi visible leur importance cachée en augmentant la fréquence globale. En d'autres termes et paradoxalement, déformer la réalité apparaît comme un bon moyen de la montrer plus correctement ! Inversement, la thlemmatisation peut également être utilisée pour exclure rapidement certains thèmes de l'analyse : parce que certaines formes ont été remplacées par le thème correspondant dans le dictionnaire, ignorer des ensembles complets de notions nécessite simplement de « désélectionner » le nom du thème auquel elles appartiennent dans la liste des formes disponibles fournie par Iramuteq (*cf. infra*). Enfin et

4. D'ailleurs, dans le cas d'Iramuteq, l'universalité des lemmes eux-mêmes est toute relative puisque la procédur de lemmatisation retenue par le logiciel s'appuie sur une liste de mots qui pose problème, notamment lorsque deux lemmes distincts sont possibles pour une même forme. Sur cette question, cf. Sarrica et al., 2016.

surtout, comme la thlemmatisation réduit considérablement le nombre de formes présentes dans l'ensemble du corpus, la sélection d'une partie de ces formes sur la base de la liste de fréquences devient plus rapide, plus clair et plus facile.

La conduite de l'ensemble de l'opération, de la thémmatisation du corpus à la sélection et à l'analyse des thèmes, repose sur un long processus d'essais-erreurs. J'ai pour ma part procédé en trois étapes. D'abord, j'ai identifié les grands thèmes portant sur les enjeux principaux de mon corpus (i.e. une liste de préoccupations comme « l'état psychologique », « les sensations corporelles », « l'économie », « l'environnement », etc.). Ensuite, j'ai établi la correspondance entre les formes du corpus et ces thèmes dans une feuille de calcul et j'ai trié les résultats en fonction des thèmes. Enfin, j'ai divisé ces thèmes en sous-thèmes dans une troisième colonne du tableur et j'ai recodé le vocabulaire en conséquence (par exemple, j'ai décliné « état_psychologique » en : anxiété, bien-être, colère, confiance, confiance, conscience, contrariété, dégoût, détresse, frustration, gêne, incertitude, intimité, méfiance, négligence, satisfaction, sécurité, surprise_curiosité). Chaque sous-thème englobe un grand nombre de formes originales (par exemple, « l'anxiété » est le sous-thème que j'ai choisi pour « alarmant », « angoisse », « anxiété », « anxieux », « appréhension », « danger », « dangereux », « effrayant », « inquiétude », « insécurité », « intimider », « menace », « nerveux », « panique », « paranoïa », « peur », « préoccupation », « risqué », « stressant », « terreur », « terrifiant », etc.). J'ai appliqué cette procédure à l'ensemble du corpus (10 756 formes), hapax exceptés (4 451 formes). Sur les 6 305 formes examinées (10 756-4 451), 2643 formes ont été thlemmatisées (41 %) selon 60 grands thèmes et 290 sous-thèmes. Seuls les sous-thèmes ont été utilisés pour la thlemmatization du corpus. Le choix d'opérer au niveau des sous-thèmes a été considéré comme un bon compromis entre recherche de lisibilité accrue et respect de la diversité lexicale. L'idée était de respecter la procédure d'analyse lexicale classique tout en simplifiant quelque peu le vocabulaire selon une simple logique de synonymie.

4. Résultats

Les développements suivants présentent les résultats de l'analyse du corpus thlemmatisé (fondée sur les sous-thèmes uniquement). Soulignons tout d'abord une différence entre les stratégies classiques et thlemmatisées, telle qu'elle ressort de la comparaison des analyses de similitude respectives des corpus lemmatisé et thlemmatisé (Fig. 1). Même avec l'approche thlemmatisée, il est toujours nécessaire de réduire le nombre de formes pour que le graphique soit lisible. Dans les exemples ci-dessous, j'ai choisi de conserver toutes les formes ayant une fréquence supérieure à 100 occurrences, tant pour la version lemmatisée que pour la version thlemmatisée du corpus. Comme on peut le voir, la version thlemmatisée permet de réduire le nombre de formes affichées pour une même fréquence choisie. La technique améliore la lisibilité et surtout la signification du graphique : loin de risquer d'être floues, dispersées ou minimisées, des notions similaires sont rassemblées sous une même rubrique. De plus, la comparaison montre clairement que, loin de se limiter à réduire le vocabulaire pour une fréquence donnée, la thlemmatisation conduit à un réseau différent.

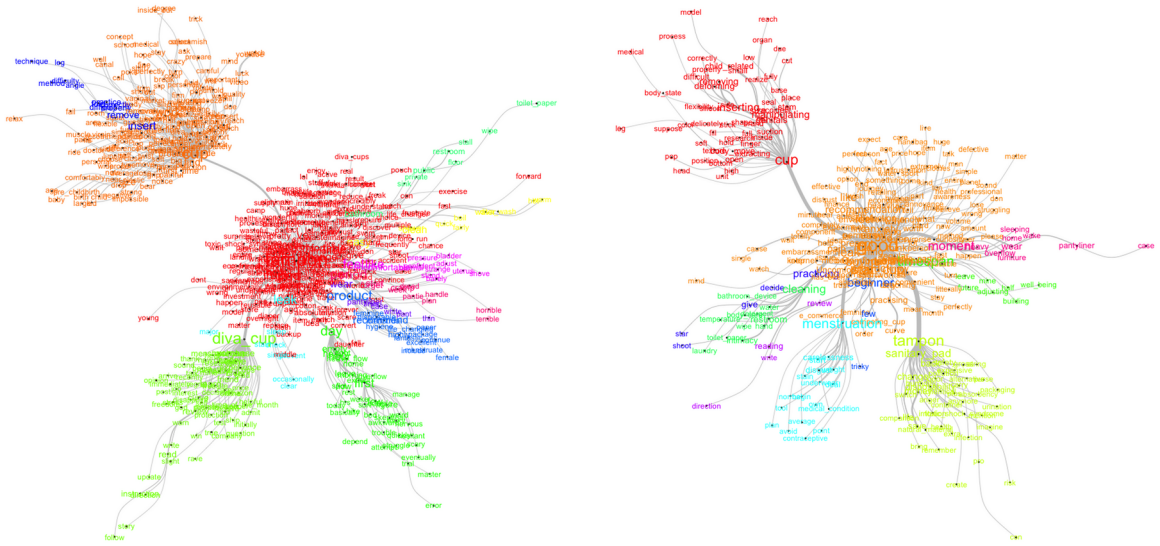

Fig. 1. Corpus lemmatisé vs. corpus thlemmatisé

Le corpus thlemmatisé étant plus lisible et mieux adapté aux objectifs de recherche que la version standard, je m'appuierai exclusivement sur lui dans les développements qui suivent. Tout d'abord, j'ai effectué une classification Reinert du corpus thlemmatisé afin d'identifier les principaux types de préoccupations que soulèvent les évaluations des consommatrices (classification simple sur segments de textes ; nombre de classes terminales requises pour la phase 1 : 8 ; nombre maximum de formes analysées : 3000 ; 14745 segments classés sur 17720 soit 83,2 %).

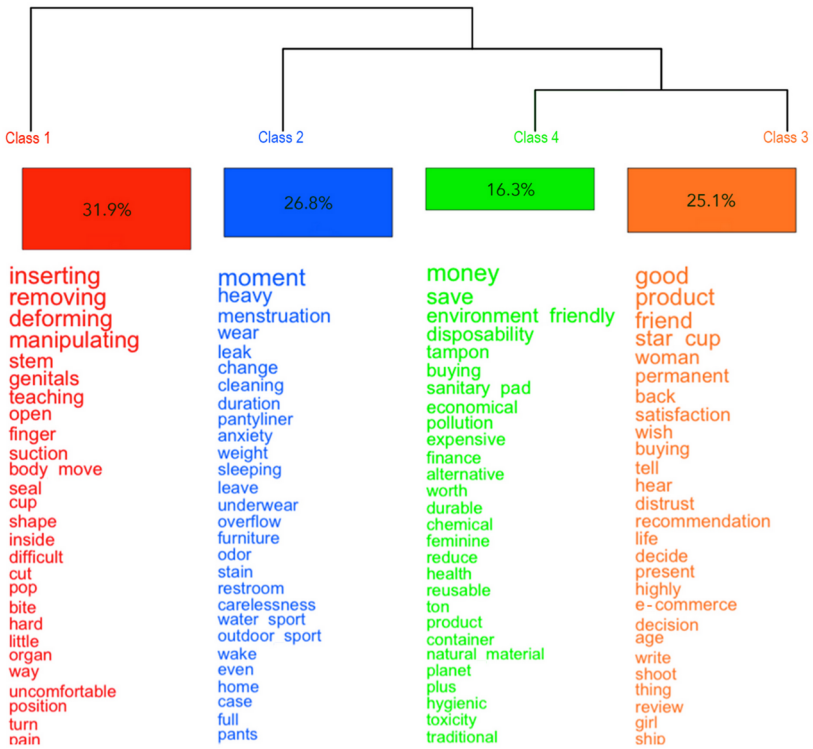

Fig. 2. Classification Reinert du corpus

Quatre classes ressortent de la classification (*cf.* fig. 2). La première classe (rouge) regroupe près d'un tiers des formes (31,9 %). Cette classe concerne la manipulation de la coupe et les problèmes afférents (l'insérer, la plier, l'enlever, savoir se servir de ses doigts, éviter les effets

de succion, etc.) Les trois autres classes correspondent à des préoccupations plus périphériques. La deuxième classe (bleue) (26,8 % des formes) concerne le contexte d'utilisation, tant en termes de temps (à quel moment, pour combien de temps, etc…) que d'espace (dans les toilettes, lors de la pratique d'un sport, à la maison…) avec l'expression des risques (« fuite ») et des sentiments associés (de l'insouciance à l'anxiété). La troisième classe (orange) (25,1 % des formes) est une classe consumériste : il s'agit d'acquérir des connaissances sur la coupe, de la juger et de partager son point de vue avec d'autres utilisatrices à travers des médias variés. La quatrième classe (verte) (16,9 %) combine les préoccupations sanitaires, économiques et environnementales, c'est-à-dire la dimension d'après laquelle le produit peut être jugé. Il est surtout intéressant de constater que les préoccupations pratiques l'emportent largement sur les autres : ce type de préoccupations appartient aux deux premières classes les plus importantes et représentent la majorité des formes (la part totale des classes 1 et 2 est de 58,7 %), tandis que les questions comme les préoccupations économiques occupent une place moindre.

Afin de mieux cerner les raisons et les raisonnements qui sous-tendent ce tableau général, je propose de me concentrer sur le sujet central des préoccupations des consommatrices. Pour atteindre cet objectif, j'ai neutralisé deux dimensions. Premièrement, j'ai exclu les tampons et les serviettes hygiéniques en désélectionnant les thèmes correspondants. En effet, si les utilisatrices de coupes comparent fréquemment leur utilisation aux alternatives classiques disponibles, le fait de retirer ces dernières de l'analyse permet de mieux se concentrer sur ce qui est dit spécifiquement sur les coupes. Deuxièmement, j'ai renoncé à l'analyse des appréciations évaluatives en excluant les jugements positifs et négatifs. Pour cela, j'ai désélectionner les thèmes « mauvais » et « bon », même s'ils représentent un nombre impressionnant de formes appréciatives et dépréciatives. En effet, ce que montre la collection de critiques généralement riche, longue et bavarde, c'est que ce qui compte le plus pour les utilisatrices, c'est moins le verdict que les expériences variées auxquelles il est lié. De même, ce qui importe le plus pour elles sont moins les critères généraux et lointains d'après lesquels les procédures d'évaluation peuvent être engagées, que l'incroyable éventail de sujets et de préoccupations qui émergent de l'expérience intime de la consommation de coupes.

J'ai effectué une analyse de similitude en conséquence. Afin d'assurer une lisibilité maximale, j'ai thlemmatisé le corpus pour les formes (lemmes et thèmes) qui apparaissent avec une fréquence de 100 ou plus (à l'exception des thèmes « serviette hygiénique », « tampon », « mauvais » et « bon », et en écartant aussi la forme « menstruation », c'est-à-dire un des thèmes les plus fréquents mais au point d'être insignifiant). J'ai conduit l'analyse de similitude pour les 299 formes restantes. J'ai exporté les données sous-jacentes vers Gephi, un logiciel de cartographie de réseaux particulièrement adapté pour traiter ces données et en améliorer la lisibilité. Les données sont directement converties au format de Gephi par Iramuteq, ce qui assure une correspondance parfaite entre les deux jeux de données. J'ai mobilisé les algorithmes « Force Atlas 2 » pour esquisser la présentation générale (un algorithme qui rapproche les nœuds liés entre eux et éloigne les autres), puis "Label adjust" et "node overlapping" pour que toutes les formes apparaissent clairement sans aucun chevauchement. Le choix de ces algorithmes de spatialisation vise essentiellement à assurer la meilleure visibilité possible. J'ai rendu la taille des étiquettes et l'épaisseur des liens proportionnels à la fréquence des formes sous-jacentes. J'ai mis en évidence les communautés fondées sur la classe de modularité (0,74), et j'ai ajusté les couleurs en conséquence (avec une couleur mixte pour les liens reliant les différentes communautés). Le résultat global est visible sur la figure 3. Les sous-thèmes thlemmatisés sont signalés par des astérisques.

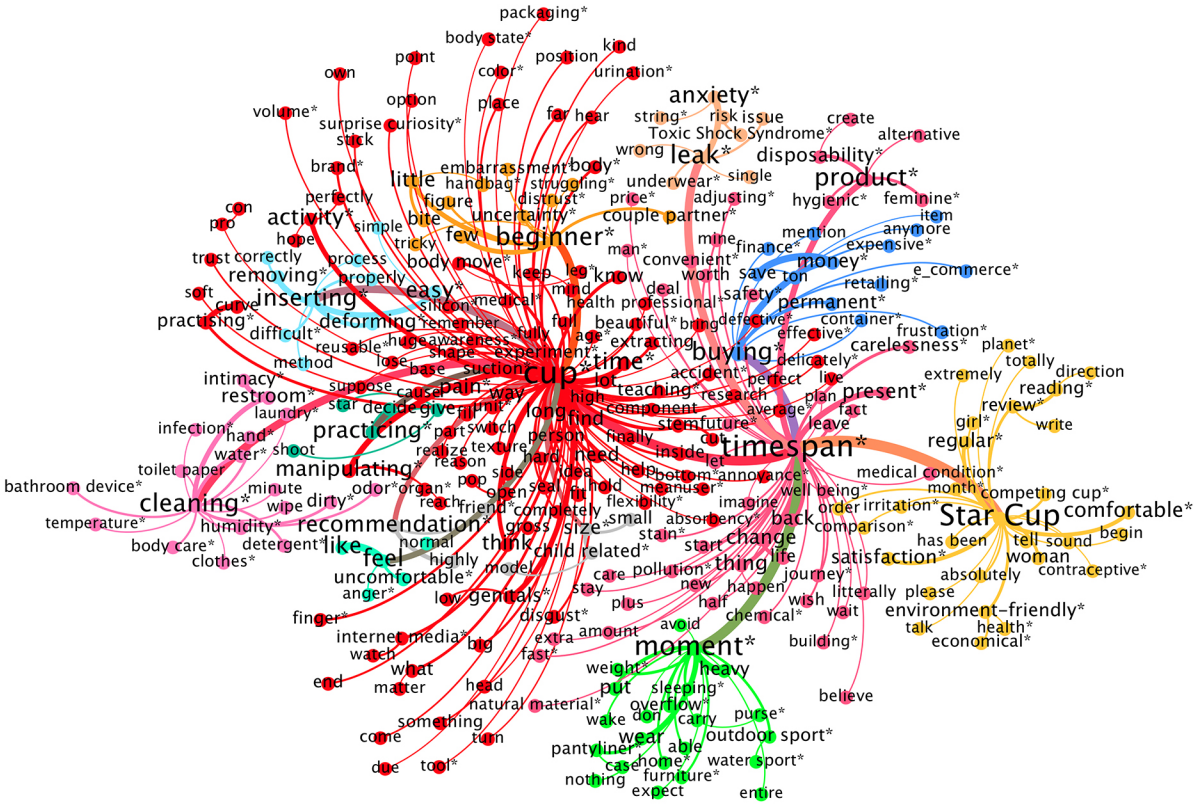

Fig. 3. Analyse de similarité du corpus de coupe thlemmatisé

Le graphique permet d'extraire au moins cinq résultats significatifs. Faute de place, je privilégie ici une lecture qualitative du graphique plutôt qu'une analyse quantitative fondée sur les propriétés structurales du réseau. Le premier résultat, et le plus frappant, est le profil bimodal de l'évocation des coupes menstruelles. En effet, comme le montre la figure 3, l'évocation du dispositif se divise en deux communautés distinctes. À droite du graphique, la coupe apparaît comme « star_cup », c'est-à-dire sous le nom de marque sur lequel portent les commentaires postés sur Amazon. Lorsqu'elle est évoqué comme telle, cette coupe est comparée à des produits concurrents (toutes les autres marques sont résumées sous le thème « competing_cup ») ; c'est un produit que l'on aborde avec des préoccupations très larges à l'esprit (« économique », « santé », « écologique ») ; c'est un produit sur lequel on recherche des informations (« lire ») ; et il fonctionne comme un sujet sur lequel on est désireux de partager des opinions personnelles (« écrire », « critiquer »). Cependant, lorsqu'elles partagent leur expérience, les utilisatrices oublient la marque, rédigent leur avis en termes plus génériques et passent à l'évocation de préoccupations plus intimes. Cette seconde approche du dispositif (« coupe ») occupe une place centrale dans le réseau et conduit à l'évocation d'un éventail de préoccupations périphériques beaucoup plus riche.

Le second résultat est que cette version centrale de la coupe est immédiatement connectée à un très grand nombre d'éléments. Il serait trop long de les commenter tous, mais on peut remarquer la forte présence du corps (« état_corporel », « doigt », « jambe », « fesses », « organes génitaux », « tête »), comme si la coupe fonctionnait comme un organe artificiel. Cette entité hybride coupe-corps nécessite à la fois l'intervention d'éléments techniques (« tige », « composant », « outil ») et le recours à une humaine (« médical », « professionnels de la santé »). La coupe est également étroitement liée à des sensations variées. Ces sensations sont à la fois physiques et psychologiques, allant de la « douleur » et de la « succion », au « dégoût » et à la « surprise_curiosité ». Enfin, il est très significatif de noter que le caractère

« réutilisable » de la coupe est mis en évidence, en fort contraste avec la lointaine évocation d'autres « produits » et leur caractère jetable (« jetabilité »).

Un troisième résultat est que, à un deuxième niveau, la coupe est reliée à une série de thèmes majeurs, semblables à des carrefours, chacun d'eux menant à des dimensions similaires. Tout d'abord, les consommatrices établissent un lien entre l'utilisation de la coupe et leur trajectoire de consommation, mettant en évidence la position particulière de la « débutante » qu'elles sont souvent, et les « incertitudes » et les préoccupations associées (« méfiance », « gêne », « lutte », « délicat »…)… au point que les débutantes ressentent rapidement le besoin de partager leurs problèmes intimes avec leur « partenaire_de_couple »). Ensuite, ces problèmes sont rassemblés autour du thème de l'« insertion ». Les opérations de manipulation complémentaires comme la « déformation » et le « retrait » appropriés du dispositif sont évidemment essentielles à cet égard, avec un soin particulier pour le savoir-faire correspondant (« méthode », « processus », « correctement », « convenablement »), et des retours d'information variés sur sa mise en œuvre (« facile » vs. « difficile »). Troisièmement, et par conséquent, les consommatrices soulignent que l'utilisation des coupes nécessite beaucoup de « pratique ». Quatrièmement, les connaissances acquises par la pratique peuvent éventuellement conduire à des attitudes de partage d'expérience (« recommandation »). Mais cinquièmement, les problèmes pratiques ne s'arrêtent pas à l'utilisation : une préoccupation importante concerne le « nettoyage » du dispositif, avec des problèmes connexes en termes de préservation de l'« intimité » dans les « toilettes » publiques, les ressources nécessaires (« appareil de salle de bain », « détergent », « eau », « papier toilette »), des problèmes sensoriels (« humidité », « odeur ») et des risques d'hygiène (« infection »).

Un quatrième résultat est l'importance du temps : le principal thème « central » lié à la coupe est la « durée », notion qui regroupe plusieurs unités de temps comme la « minute », l'« heure », le « jour », la « semaine », etc. Cela a bien sûr un sens pour la consommation d'un produit lié aux menstruations. Plus intéressante toutefois est la présence de préoccupations plus spécifiques liées à des durées différentes. Certaines de ces préoccupations concernent des événements inattendus, comme les « taches » et d'autres types d'« accidents ». De manière significative, une préoccupation périphérique majeure concerne le risque de « fuite(s) » et l'« anxiété » qui l'accompagne. Plus important encore, la « durée » est fortement liée à différents « moment(s) », un thème qui qualifie les moments particuliers où des événements sont susceptibles de se produire (« matin », « midi », « après-midi », « soir », « nuit »…). Comme on peut le constater, le temps est lié à l'espace : le souci des moments varie selon les lieux et les activités qui s'y rapportent (quand on « dort » ou quand on est « chez soi ; quand on pratique un « sport_de_plein air » ou un « sport_de_eau »). On découvre le caractère réticulaire de l'utilisation de la coupe qui fait partie d'un agencement beaucoup plus large que les composants du produit lui-même. Enfin, la durée concerne également le temps qu'il faut non seulement pour utiliser le produit, mais aussi pour l'acheter, avec le souci relatif de l'« argent » que l'on dépense, des canaux de distribution auxquels on recourt (« vente au détail » vs. « commerce électronique »), etc.

Enfin et surtout, un cinquième résultat est que les questions de santé et d'environnement, même « boostées » par la procédure de thlemmatisation, apparaissent comme des préoccupations quelque peu perdues parmi les nombreux autres soucis plus terre-à-terre et pratiques que je viens de passer en revue. En résumé, l'analyse des similitudes du corpus thlemmatisé montre clairement que la consommation des coupes est d'abord une question très pratique et matérielle, une « affaire d'usage », et cette dimension pragmatique est évidemment liée à un vaste réseau de préoccupations interdépendantes. En d'autres termes, si le comportement d'achat s'opère souvent en référence à divers « ordres de valeur » externes (comme le prix, le commerce équitable, la durabilité, la responsabilité sociale des entreprises,

etc.), le comportement de consommation s'opère plutôt en relation avec une infinité d'« ordres de préoccupation », y compris tous les précédents, sans oublier des questions plus pratiques et intimes comme le confort, l'intimité, l'anxiété, etc.

5. Conclusion

Mon propos a consisté à présenter une astuce destinée à faciliter la lecture des graphiques produits lors d'analyses de similitude par Iramuteq en aidant le logiciel à identifier les unités thématiques qui rassemblent des séries de mots disjoints. La démarche consiste à circonscrire les champs lexicaux sur lesquels porte prioritairement la recherche, et à substituer ces champs lexicaux, ou thèmes, aux lemmes du dictionnaire. L'opération, même conduite de façon largement inductive en privilégiant un grand nombre de sous-thèmes plutôt qu'un petit nombre de thèmes, réduit significativement le nombre de formes affichées pour une fréquence donnée, et renforce aussi la mise au jour des liens entre ces unités thématiques et les mots associés. Le retraitement du réseau sous Gephi permet d'accroître encore la lisibilité, notamment grâce à l'usage d'algorithmes évitant les chevauchements. Les limites de cette méthode résident dans son caractère arbitraire (la mise au point des catégories pertinentes est fastidieuse et nécessite d'être mûrement réfléchie et argumentée pour être acceptable) et surtout dans son aspect chronophage : le recodage du vocabulaire est un processus très long, ce qui entre en contradiction avec les gains de temps que l'on recherche lorsqu'on emploie des méthodes d'analyse automatique des données. Toutefois, le gain de lisibilité peut justifier, au moins dans certains cas, le recours à la thlemmatisation, dès lors que l'on souhaite faire en sorte que le chercheur et ses lecteurs puissent mieux partager la lecture des données.

Références

Beuscart J.-S., Mellet K. and Trespeuch M. (2016). Reactivity without legitimacy? Online consumer reviews in the restaurant industry, *Journal of Cultural Economy*, 9(5): 458-475.

Blank G. (2006). *Critics, Ratings, and Society: The Sociology Of Reviews*. Lanham, Rowman & Littlefield Publishers.

Cochoy F. (2021). Patents as vehicles of social and moral concerns: The case of Johnson & Johnson disposable feminine hygiene products (1925–2012), *Science, Technology & Human Values*, first published online, February 10, https://doi.org/10.1177/0162243921992861.

Loubère, L. et Marchand, P. (2019). Les thématiques du Grand Débat et du Vrai Débat : approche textométrique. JE MetSem et Mate-Shs, 5 et 6 décembre, Sciences Po Paris.

Liu B. (2012). *Sentiment analysis and opinion mining*. San Rafael, CA: Morgan & Claypool.

Marchand P. and Ratinaud P. (2012). L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française, Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelle, pp. 687–699.

Petty R. E., Cacioppo J.T. and Schumann D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, 10(2): 135-46.

Ratinaud P. and Marchand P. (2015). From lexical scopes to social representations. A preliminary thematic approach to the National Assembly debates (1998-2014). *Mots. Les langages du politique*, 108: 57-77.

Sarrica M., Mingo I., Mazzara B. and Leone G. (2016). The effects of lemmatization on textual analysis conducted with IRaMuTeQ: results in comparison. 13ème Journées internationales d'Analyse statistique des Données Textuelles.