ATELIER EPISTHEMETH
"Analyse de données textuelles"

Iramuteq : l'usage exploratoire de l'analyse lexicale de données textuelles

Partage d'expérience par Elzbieta Sanojca

Atelier EPISTHEMETH organisée par l'équipe PRT5 du CREAD (EA 3875)
Université Rennes 2, 26 janvier 2024

Explorer, vérifier, densifier : modalités d'usage

Exemples :

- Compétences collaboratives en formation des adultes (thèse, 2018)
- Séminaire 'AgoraPro' un espace de productions de savoirs partagés : analyse exploratoire (COM, 2023)
- Les réseaux professionnels de masseur-kinésithérapeute , usages de Facebook pour apprendre : étude exploratoire (ACTI, 2021)

… modestes : en analyse « secondaire »

Exemple 1 : Compétences collaboratives en formation des adultes (thèse, 2018)

Usage présenté « seulement » comme une justification méthodologique

(dans : Méthodologie / Méthodes de traitement des données et clés d'analyse : construire le sens)

- Inspiration de la procédure *Knowledge Discovery in Databases (KDD) :*
  - développer les outils nécessaires pour contrôler un grand volume d'informations.
  - la procédure suit plusieurs étapes : collecte des données, sélection des données pertinentes, prétraitement des données (*preprocessing*), , transformation, recherche et interprétation de significations, interprétation puis évaluation des modèles (Fayad *et al.*, 1996).

- Double traitement en phase « prétraitement » : *top down* avec R-based Qualitative Data Analysis (RQDA) et *bottom-up* avec Iramuteq, équivalent en logiciel libre d'Alceste.

Exemple 1 : Compétences collaboratives en formation des adultes (thèse, 2018)

Ce qui en est dit : (extrait)

Après avoir nettoyé les 36 entretiens intégralement retranscrits des questions de l'enquêtrice et retirés les signes non valides, le corpus comportait 278 074 occurrences. Ce corpus a ensuite été subdivisé en trois thématiques : « faire collaboratif », « apprentissage », « réinvestissement », pour tenter d'en déduire, pour chacune d'entre elle, des classes lexicales, puis la proximité et les relations entre les éléments (analyse de similitude). Le résultat escompté s'est montré décevant, car basé sur la seule fréquence, coupée de l'analyse du sens ; les occurrences les plus fréquentes identifiées semblaient indiquer, non des spécificités collaboratives de l'activité, mais plutôt des dominantes langagières. Jugés inutilisables pour l'étape suivante de l'analyse, les résultats obtenus ont été écartés de l'étude. Seule l'attention portée à la fréquence d'occurrence « outil » a été retenue comme une vigilance à avoir dans l'analyse thématique (*top-down*), qui semblait en définitive mieux adaptée. À titre illustratif, l'exemple de l'analyse des similitudes du sous-corpus thématique « faire collaboratif » est présenté ci-dessous (figure 19).

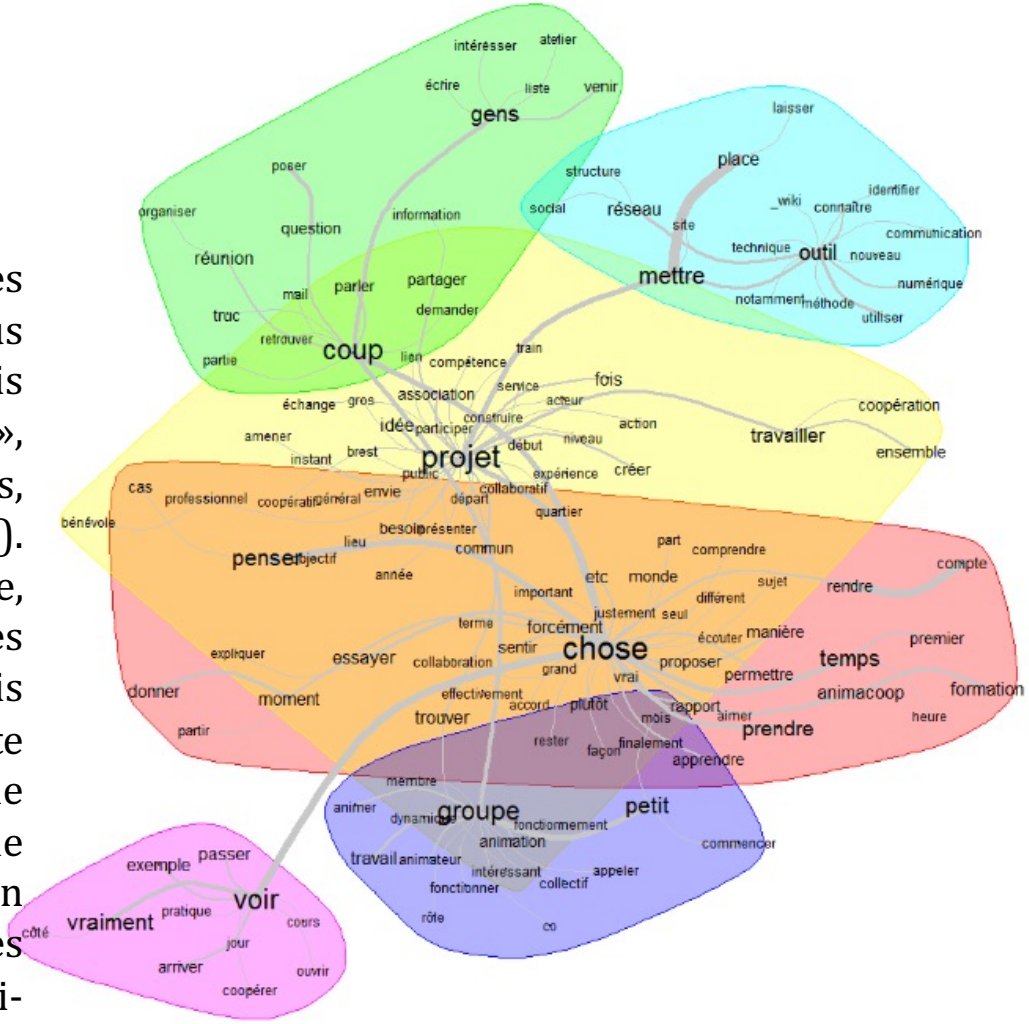


Figure 19 : Exemple du traitement (*bottom-up*) via Iramuteq

Exemple 1 : Compétences collaboratives en formation des adultes (thèse, 2018)

Les plus méthodologiques

- Analyse systématisée d'un corpus important :
  36 entretiens (278 074 occurrences)
- Analyse dit *bottom-up* en complément de *top down*
- Analyse de fréquences seule, jugée insuffisante
- Singularité objectivée « effet 'tiens ' »

Les plus personnels

- Apprentissage du vocabulaire (AFC , AS –
- Apprentissage de la « technique » de l'analyse lexicale (niveau « je sais à peu près »)
- Manque : justification des calculs

En résumé : les fonctions « explorer », « vérifier » présentées sous forme d'une justification méthodologique

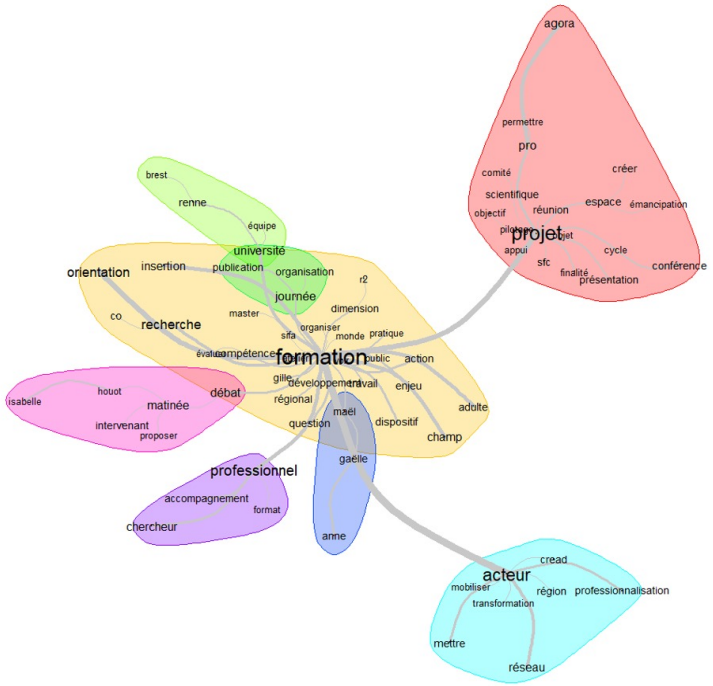
Exemple 2 : ‘AgoraPro’ un espace de productions de savoirs partagés : analyse exploratoire (COM, 2023, RUMEF)

Fonction « exploration » pure

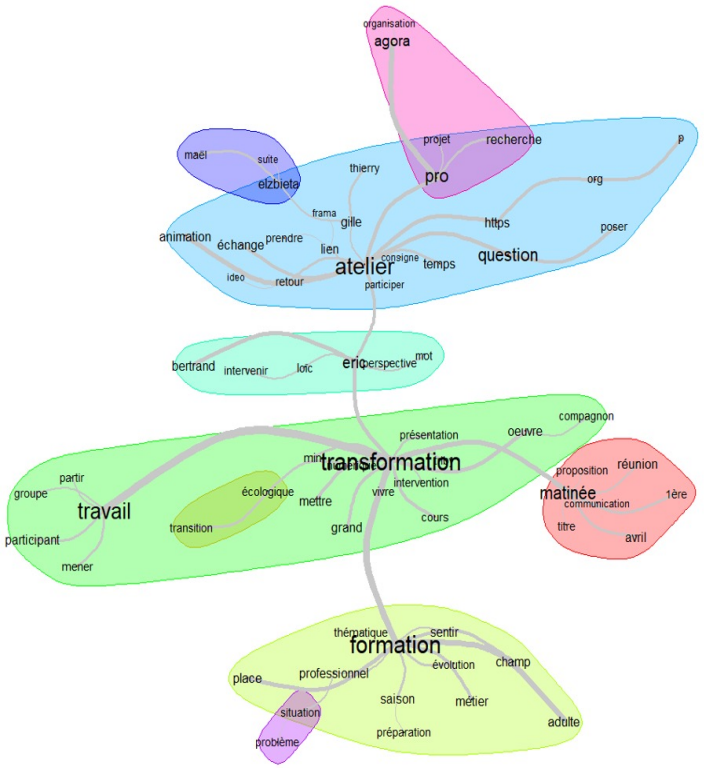
- Données textuelles : documents de présentation du projet, comptes-rendus des réunions produit en trois ans
- 14 364 occurrences (relativement peu)
- Présentation en communication lors d’un colloque RUMEF (sans ACTI)

Exemple 2 : ‘AgoraPro’ un espace de productions de savoirs partagés : analyse exploratoire (COM, 2023)

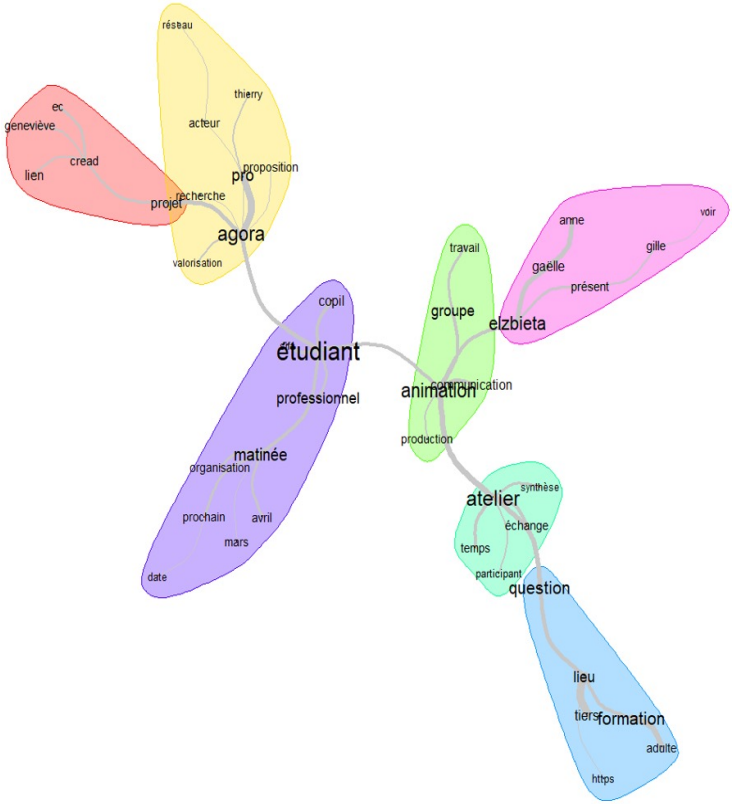
Ce qui est montré et pour en dire quoi ?



Année 1



Année 2



Année 3

Exemple 2 : ‘AgoraPro’ un espace de productions de savoirs partagés : analyse exploratoire (COM, 2023)

Les plus méthodologiques

- Valorisation des données « secondaires » (Comptes-rendus , documents de présentation, etc.)
- Analyse longitudinale « potentiellement » efficace

Les plus personnels

- Intérêt pour les données du type productions des personnes/groupes « traces de l’activité »

En résumée : les fonctions « explorer » appliquée aux données textuelles secondaires (productions d’un groupe)

Exemple 3 : L'usage de Facebook pour apprendre chez masseurs-kinésithérapeutes : étude exploratoire (ACTI, 2021)

Fonction « exploration » intégrée dans une étude doctorale

Pour l'étude de ce phénomène, l'attention s'est portée dans cette communication sur la valorisation de la formation non formelle et informelle des masseurs-kinésithérapeutes (MK). Comment ces activités participent à l'actualisation de leurs connaissances professionnelles ?

Le corpus étudié se compose de 105 "posts" *Facebook* envoyés entre mai 2020 et avril 2021, issus de 26 groupes parmi les 45 observés. Les posts retenus comportaient soit une demande d'aide formulée par un MK, soit des échanges en réaction à un post traitant d'un aspect du métier

Justification d'usage :

« […] rechercher des spécificités ou des invariants d'un texte (…), (chercher) *à identifier, dans les textes examinés, différentes classes regroupant chacun des mots fréquemment employés ensemble dans certaines phrases, mais peu présents dans les autres », « à mettre en évidence et des caractéristiques des thématiques spécifiques »* et a *« servir de base a des cartographiques des thèmes et des contenus abordes par les auteurs »* (O. Las Vergnas, 2021, cité Milanovic et a., 2021).

Exemple 3 : L'usage de Facebook pour apprendre chez masseurs-kinésithérapeutes : étude exploratoire (ACTI, 2021)

Fonction « exploration » intégrée dans une étude doctorale

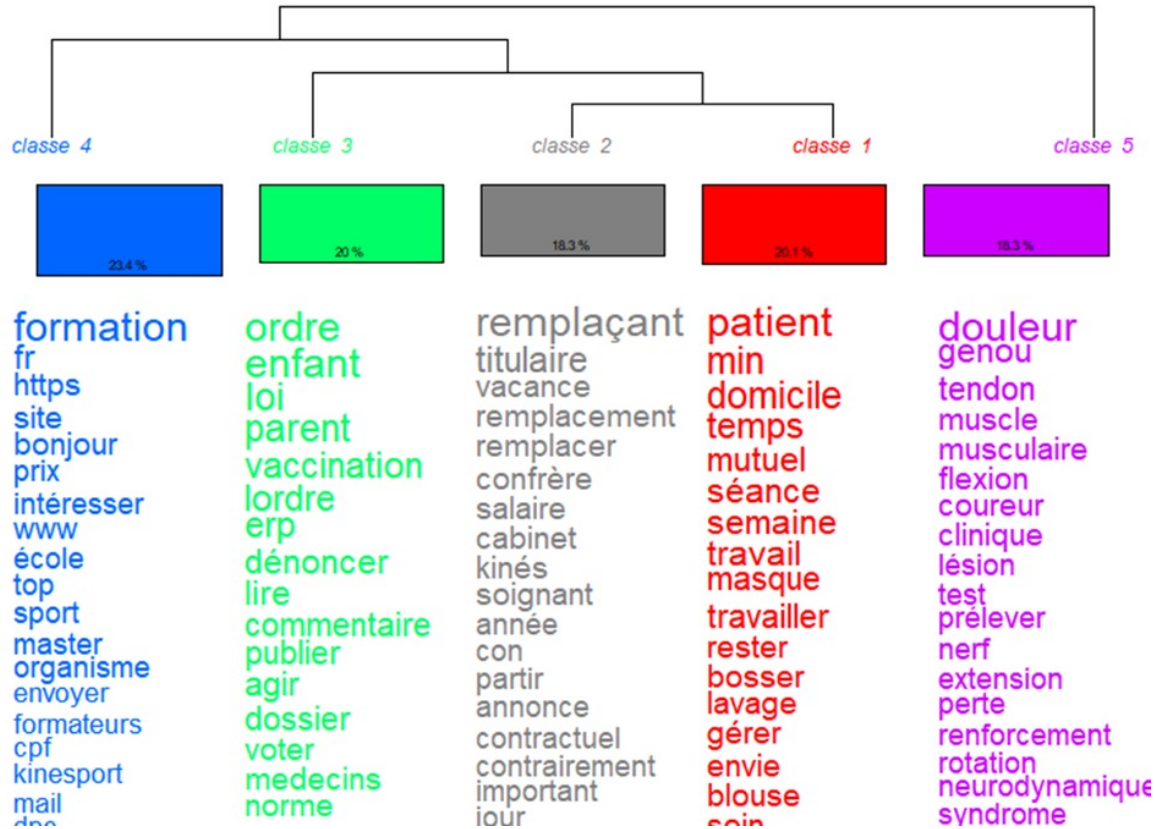


Figure 2 : Classification en 5 classes de lexiques employés dans les posts Facebook relevés.

Classification : analyse de fréquences

Interprétation

-« ...en ouvrant le concordancier (outil d'*Iramuteq*® permettant de replacer les mots dans leurs posts) nous avons pu les remettre en contexte » (Milanovic, et al., 2021)

Exemple 3 : L'usage de Facebook pour apprendre chez masseurs-kinésithérapeutes : étude exploratoire (ACTI, 2021)

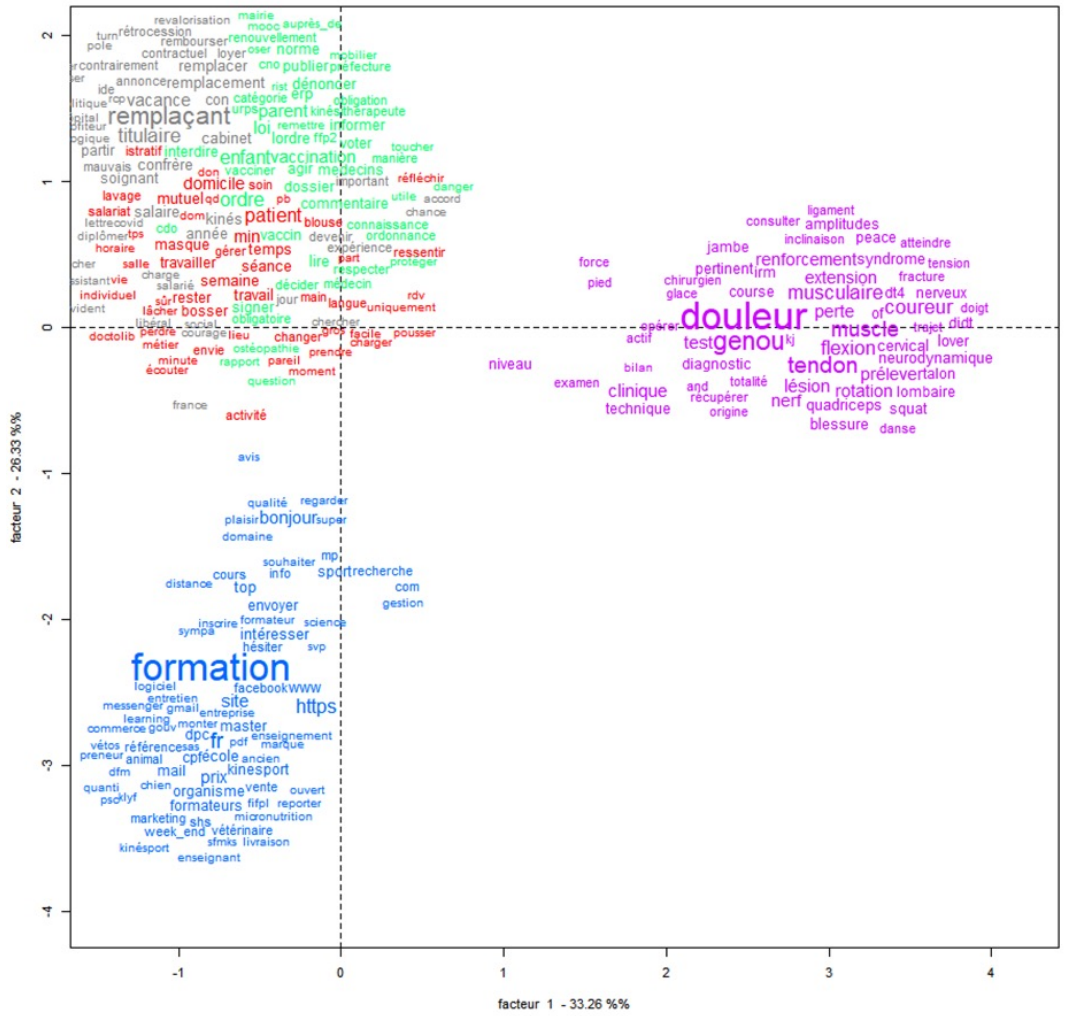


Figure 3 : analyse factorielle du corpus

Une analyse factorielle de ces classes (AFC)

Interprétation :

Nous remarquons dans le graphique de l'analyse factorielle ci-dessus que les questions concernant la formation sont à l'opposé des questions traitant des douleurs. Cela pourrait s'expliquer par la valeur accordée à la formation où la priorité serait la qualité de l'organisme et de la formation plutôt que son rapport avec la douleur du patient. Les questions et réponses évoquant l'ordre des MK sont entremêlées avec les problématiques autour des patients et des questions de remplacements. Cette derrière information peut s'expliquer par le fait que de l'ordre des MK réalise de nombreuses préconisations autour du traitement des patients, des devoirs des MK, et assiste les MK dans leur résolution de conflit entre MK. (Milanovic, et al., 2021)(

Exemple 3 : L'usage de Facebook pour apprendre chez masseurs-kinésithérapeutes : étude exploratoire (ACTI, 2021)

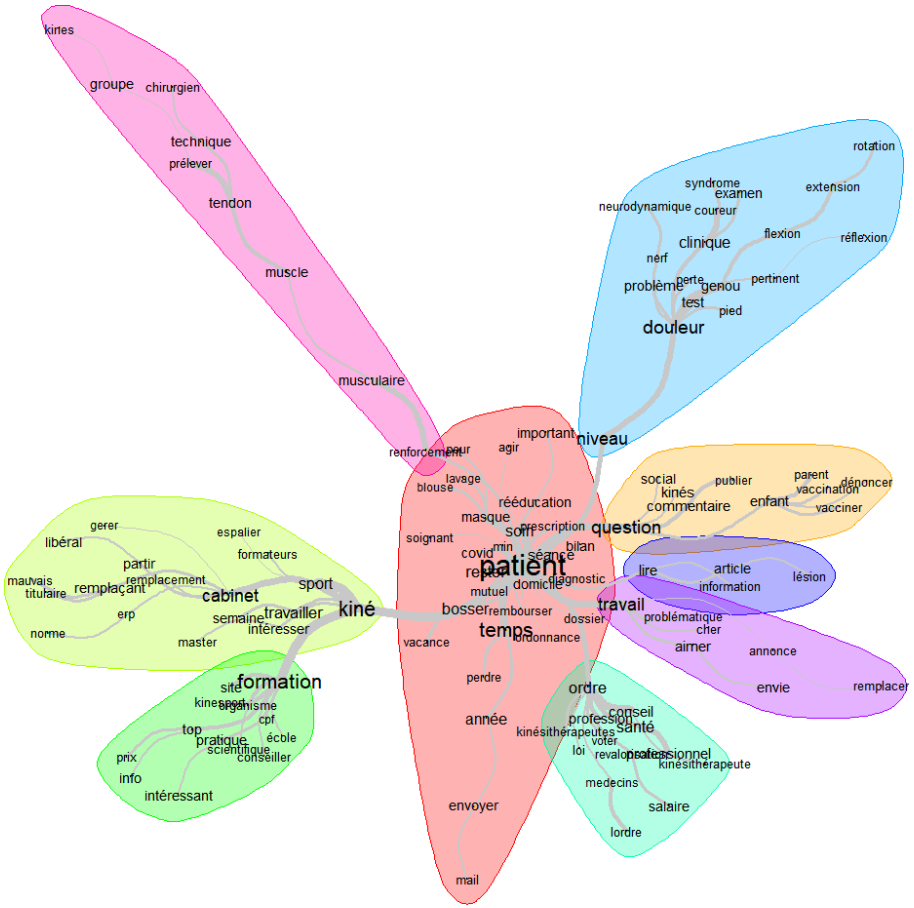


Figure 4 : Analyse des similitudes

Une analyse factorielle de ces classes (AFC)

Interprétation descriptive :

Afin de comprendre comment ces thèmes sont reliés entre eux, nous avons réalisé une analyse des similitudes dans notre corpus représenté ci-dessous (figure 4)

Ce que l'on voit est que le patient est au centre du questionnement des MK….. Nous voyons aussi que les discussions évoquant la douleur (…..) Ce schéma met également en évidence l'intérêt des MK pour les techniques de renforcement musculaire……

Exemple 3 : L'usage de Facebook pour apprendre chez masseurs-kinésithérapeutes : étude exploratoire (ACTI, 2021)

Les plus méthodologiques

Analyse descriptives des données exploratoires pour affiner les questions de recherche

Complémentarité d'analyse par l'articulation des types d'analyse (Classes + AFC + AS)

Les plus personnels

- Possibilité d'exploitation des analyses « Iramuteq » en communication scientifique
- (à vérifier) distinction de qualité de données textuelles pout analyse lexicale : entre les données 'texte' et données 'verbatim' transcrit

En résumé : la fonction « explorer plus» formalisée par la description interprétative des données exploratoires

ANNEXES

Vocabulaire Iramuteq

Le principe de fonctionnement

Son fonctionnement se différencie des fonctionnalités de RQDA puisqu'il est fondé sur les proximités entre les mots employés et la statistique fréquentielle. Comme le décrivent Fallery et Rodhain (2007), ce logiciel procède dans un premier temps à la fabrication d'un lexique de mots, puis à un découpage du texte en unités, pour en construire ensuite une matrice de présence/absence des mots en lien avec l'unité de texte. À cette matrice sont alors appliquées les méthodes de l'analyse des données multidimensionnelles telles que l'analyse factorielle de correspondances ou l'analyse des similitudes. Il en résulte des classes, des catégories ou des oppositions.

Vocabulaire Iramuteq*

Les formes textuelles

Les formes textuelles (ou formes simples) sont des suites de signes séparés par des blancs (ou d'autres caractères comme les points, les virgules...), c'est-à-dire, en général, des mots. Elles subissent dans Alceste une lemmatisation qui les remplace par leur forme réduite : ainsi, une forme verbale est réduite en infinitif (mangerai devient manger), un substantif pluriel est réduit en singulier (arbres devient arbre), une forme élidée est réduite sans élision (l' devient le).

Les unités de contexte élémentaires

Le logiciel découpe le texte à étudier en unités de contexte élémentaires (u.c.e.), ou segments, de taille réduite

Ces u.c.e. sont composées d'une ou plusieurs lignes de texte consécutives d'environ 200 caractères et terminées si possible par une ponctuation, sinon par un séparateur comme un blanc. Elles sont regroupées par concaténation en unités de contexte (u.c.) de telle sorte que ces u.c. contiennent un nombre minimal de formes analysables différentes. Ce nombre est calculé pour optimiser la stabilité des classifications.

* Cf Loubère, L., Ratinaud, P. (2014). Documentation IRaMuTeQ 0.6 alpha 3 version 0.1 (en ligne) http://www.iramuteq.org/documentation/fichiers/documentation_19_02_2014.pdf

Vocabulaire Iramuteq

Les mondes lexicaux (classes)

Une classification descendante hiérarchique (…) regroupe ces unités de contexte en classes, ou mondes lexicaux(…) différencies par la distribution de leur vocabulaire

La classification descendante hiérarchique est particulièrement adaptée pour des tableaux comportant une très forte proportion de zéros. En effet, le tableau lexical comporte la valeur 1 quand la forme lemmatisée est présente dans l'u.c., et la valeur 0 lorsqu'elle est absente. Le tableau est réorganisé pour produire deux classes de formes les plus contrastées possibles (c'est-à-dire employées à des moments distincts) ; la plus grande est ensuite découpée en deux, etc. On réalise ainsi dix classes ; ce nombre entant « élevé », les classes risquent de dépendre beaucoup du découpage en u.c. On cherche donc le nombre de classes stables, en réalisant deux classifications successives, sur des u.c. de dimension légèrement différente, et en les comparant (par croisement et utilisation d'un test du chi2). On restreint ensuite les classes aux formes qui sont présentes dans les deux classifications. On obtient ainsi des classes d'énonces significatifs qui renvoient à des mondes lexicaux. (…) « Dans la pratique, le nombre de classes n'a que peu de signification, ce qui est important c'est la forme de l'arbre de classification et la stabilité des classes obtenues »).

Références :

Sur l'analyse lexicale (dont Iramuteq)

- Fallery, B. & Rodhain, F. (2007). Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique. *XVI*ème *Conférence Internationale de Management Stratégique.* AIMS, Montréal, 6-9 juin 2007. Repéré à : https://hal.archives-ouvertes.fr/hal-00821448
- Marpsat, M. (2010) « La méthode Alceste », Sociologie [En ligne], N°1, vol. 1 URL : http://journals.openedition.org /sociologie/312
- Las Vergnas, O. (2021) Le numérique : entre démystification, modélisations et retours d'expérience, *Éducation permanente* 226. (à vérifier son usage d'iramuteq)
- (Milanovic, Y., Sanojca, E. et Triby, E. (2021). Les réseaux professionnels en ligne : quels savoirs
- y sont en jeu ?. Biennale internationale de l'éducation, de la formation et des pratiques professionnelles.
- Faire/Se faire 2021, Sep 2021, Paris, France. hhal-0373129

Documentation pour utiliser Iramuteq

- Loubère, L., Ratinaud, P. (2014). Documentation IRaMuTeQ 0.6 alpha 3 version 0.1 (en ligne) http://www.iramuteq.org/documentation/fichiers/documentation_19_02_2014.pdf
- Site : www.iramuteq.org/
- Vidéo de présentation par le concepteur d'Iramuteq_Pierre Ratinaud

Exemple d'usage

Smyrnaios, N., Bousquet, F. et Marty, E. (2016). La mobilisation en ligne contre la Loi travail: enquête sur les réseaux et les discours (en ligne)